Applications of Leveling Methods to Properties of Small Molecules and Protein Systems

Leherte, Laurence

Published in:
Innovations in Computational Chemistry

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

CHAPTER 8

APPLICATIONS OF LEVELING METHODS TO PROPERTIES OF SMALL MOLECULES AND PROTEIN SYSTEMS

LAURENCE LEHERTE

Laboratory of Computational Physical Chemistry, Unit of Theoretical and Structural Physico-Chemistry, Department of Chemistry, Namur MEdicine and Drug Innovation Center (NAMEDIC), University of Namur, Rue de Bruxelles 61, B-5000 Namur, Belgium,
Tel.: +32-81-72-45-60, E-mail: laurence.leherte@unamur.be

CONTENTS

ABSTRACT

Despite the advent of high performance computing resources, the calculations applied to the large systems may remain intractable. Methods to reduce the level of details are therefore essential to allow fast calculations, but also to provide new insights into the systems under study. In this chapter, various

techniques and application domains related to the leveling of molecular properties through low-resolution, smoothing, denoising, or coarse-graining approaches, are presented. A focus is done on Gaussian smoothing, wavelet multi-resolution analysis, crystallography-based methods, as well as discretization methods. An emphasis is given on the use of smoothed charge density distribution functions and their extrema to generate reduced point charge models (RPCM) of proteins. Molecular dynamics simulations based on RPCMs are reported for three ubiquitin complexes. Results are discussed based on the ability of such models to generate stable protein-ligand conformations.

8.1 INTRODUCTION

Computer resources have now become sufficiently powerful to enable simulations of large systems at a classical level. Therefore, biological macromolecules and supramolecular complexes, for example, can be modeled with atomic details. However, when the systems include huge numbers of degrees of freedom and/or environment considerations, or when they contain unnecessary details like noise, calculations may remain too long. Low-resolution and smoothing techniques can thus bring an aid to the modeling of large systems. On the experimental point of view, low-resolution representations are also extremely useful in the refinement or interpretation of images generated by experimental approaches such as electron microscopy (EM) or X-ray diffraction. For example, a challenge in structural biology is to establish the structure of complex systems, which require high-resolution structural determination methods to generate atomic models of the individual components. Complexes created from the well-resolved individual components are imaged at a lower resolution to validate the interpretation of the experimental low-resolution image.

In this chapter, some methods that are used to level out molecular properties and fields, or to generate reduced discrete molecular representations of biomolecules, are presented. Applications are then given in various fields, with an emphasis on the use of smoothed charge density (CD) distribution functions and their extrema to generate reduced point charge models (RPCM) of proteins. Specific calculations based on RPCMs are reported for ubiquitin-ligand complexes modeled through Molecular Dynamics (MD) simulations. Results are discussed based on the ability of such models to generate stable protein-ligand conformations.

8.2 METHODS

8.2.1 SPLINE APPROXIMATION

The approximation of mathematical functions within a given interval defined by control points is among the numerous applications of, e.g., B-splines (basis splines) in science. Such piecewise polynomial functions (Eq. 1), whose shape is determined by the control points, are characterized by continuity conditions at their junctions, and are for example often used to smooth experimental data.

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_{k-1} x^{k-1} \quad (1)$$

In Eq. (1), k and $k-1$ are the order and the maximum degree of the B-spline function, respectively. The resolution of the smoothed function thus depends on the number of polynomials used to approximate the initial function and on their order k. The use of B-splines does not require any a priori knowledge regarding the trend followed by the data. In a relatively recent paper, Klasson details how to construct spline functions in spreadsheets to smooth experimental data [1]. Earlier, Oberlin and Scheraga [2] used B-spline functions to approximate, through a pre-calculated potential energy, the interaction energy between rigid and fixed parts of a molecular system. A well-known application of splines is the ribbon representation of molecules like proteins and DNA, which allows a clear picture of the secondary structure and fold of the macromolecules (Figure 8.1). Contributions to such representations were brought by Carson [4] who also used B-splines to model molecular surfaces [5]. In his work, Carson applied B-spline filters to represent protein backbones, folds, as well as surfaces, with a suggested implication in structure-based drug design. Additional references regarding the approximation of molecular surfaces can be found in the work of Bajaj et al. [6].

8.2.2 GAUSSIAN TRANSFORMATION

Gaussian transformation, also known as Gaussian smoothing or blurring, belongs to the so-called kernel-based techniques. A function *f(x)* is smoothed through a convolution product with a Gaussian smoothing kernel *G(x-y,t)*:

$$F(x,t) = \int G(x-y,t) f(x) dx \quad (2)$$

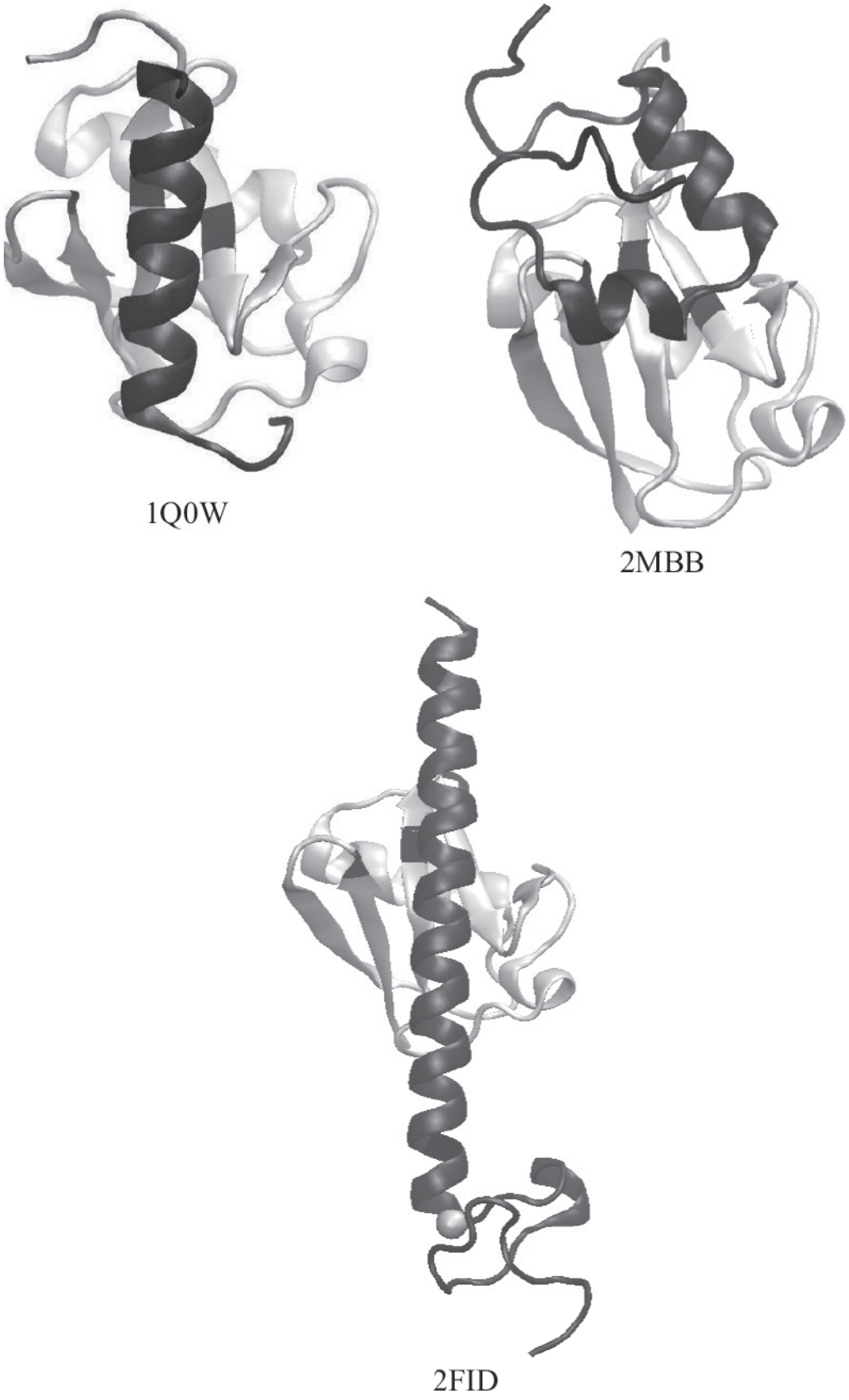


FIGURE 8.1 Catmull-Rom spline representation of Ubiquitin complexes obtained using VMD [3]. The ligand is displayed in black. Residues Leu8, Ile44, and Val70 of Ubiquitin are shown in black. The zinc ion in structure 2FID is shown as a gray sphere.

where:

$$G(x-y,t)=\frac{1}{2\sqrt{\pi t}}e^{-\frac{(x-y)^2}{4t}} \quad (3)$$

The convolution product leads to mathematical equations that involve a smoothing parameter *t*, also called deformation or smoothing parameter, that is modulated to smooth (*t* is increased) or unsmooth (*t* is decreased) *f(x)*. Various transformations of elementary functions are reported by Moré et al. [7]. Such a smoothing technique is easily applicable to three-dimensional (3D) molecular properties represented themselves by Gaussian functions. Indeed, the convolution products can be calculated using analytical formula as illustrated later in the paper for the treatment of electron density (ED) and CD fields. Smoothing the function *f(x)* through a convolution product with a Gaussian is equivalent to define *f(x,t)* as its deformed version that is directly expressed as the solution of the diffusion equation according to the formalism presented by Kostrowicki et al. [8]. The method is thus known as diffusion equation method (DEM). In their paper, the authors used the procedure to the smoothing of interaction potentials in order to facilitate the global optimization of atom clusters. The technique was also applied to the prediction of a crystal structures, like S_6 [9]. An example of a one-dimensional (1D) smoothed potential energy function *f(x,t)* is illustrated in Figure 8.2. As the smoothing factor *t* increases, the two initial potential energy wells progressively disappear to eventually lead to a single minima.

In the following sub-sections, 3D molecular fields like ED and CD are given as particular application cases.

8.2.2.1 Application to Promolecular Electron Density Distribution Functions

Promolecular models, i.e., molecular models built with non-interacting atoms, have often turned out to lead to very good approximated representations of ED distributions for the purpose of a number of applications as varied as chemical bond analysis or molecular similarity applications [10–17]. In the promolecular atomic shell approximation (PASA) approach developed by Carbó-Dorca and co-workers, a promolecular ED distribution r_M is calculated as a weighted summation over atomic ED distributions r_i, i.e., $\rho_M = \sum_i^{No.atoms} Z_i \rho_i$, where Z_i is the atomic number of atom *i*. r_i is described in

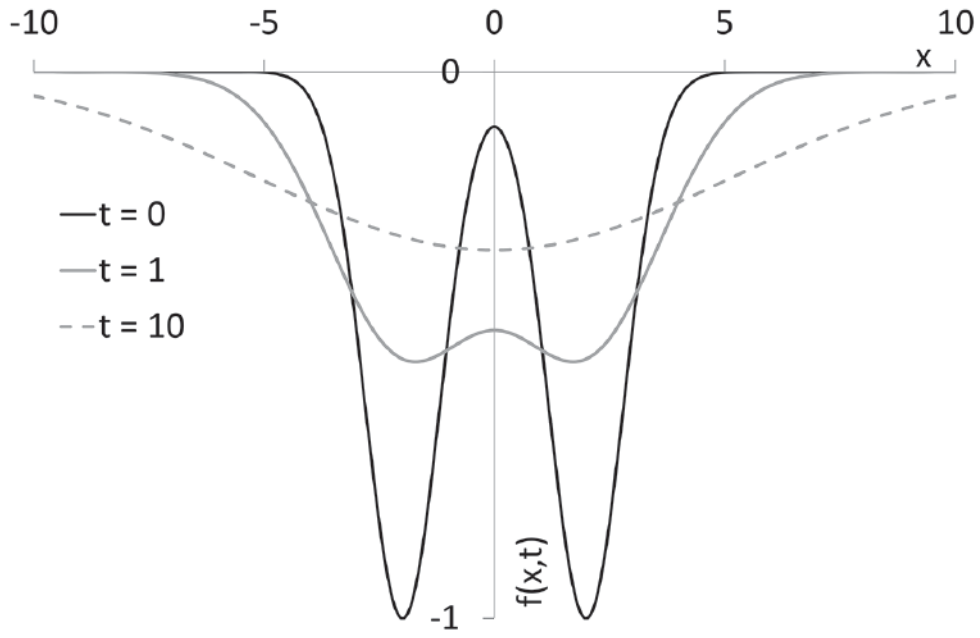


FIGURE 8.2 Gaussian smoothing of a hypothetical 1D potential energy function $f(x,t) = -\left(\left(1+4b\right)^{-1/2} e^{-b*(x-2)^2/(1+4b)} + \left(1+4b\right)^{-1/2} e^{-b*(x+2)^2/(1+4b)}\right)$ with $b = 0.75$ (arbitrary units). The original unsmoothed signal is obtained when $t = 0$.

terms of series of squared 1s-type Gaussian functions fitted to atomic basis set representations [18,19]:

$$\rho_i(\mathbf{r}-\mathbf{R}_i) = \sum_{j=1}^{5} w_{i,j}\left[\left(\frac{2\varsigma_{i,j}}{\pi}\right)^{3/4} e^{-\varsigma_{i,j}|\mathbf{r}-\mathbf{R}_i|^2}\right]^2 \tag{4}$$

where $\mathbf{R}_i$ is the position vector of atom i, and $w_{i,j}$ and $z_{i,j}$ are the fitted parameters, respectively. The number of 1s-type functions used to approximate the ED of an atom may vary depending on the model. When applied to r_i as given in Eq. (4), the Gaussian smoothing approach leads to:

$$\rho_{i,t}(\mathbf{r}-\mathbf{R}_i) = \sum_{j=1}^{5} a_{i,j}(1+4b_{i,j}t)^{-3/2} e^{\frac{-b_{i,j}|\mathbf{r}-\mathbf{R}_i|^2}{1+4b_{i,j}t}} \tag{5}$$

where:

$$b_{i,j} = 2\varsigma_{i,j} \qquad a_{i,j} = w_{i,j}\left(\frac{b_{i,j}}{\delta}\right)^{6/4} \tag{6}$$

In this context, the smoothing parameter t is seen as the product of a diffusion coefficient with time. Figure 8.3 shows the evolution of the promolecular ED distribution of Piroxicam, an anti-inflammatory drug molecule, as t increases from 0.0 to 2.5 bohr^2. Coordinates were retrieved from the crystallographic structural database (CSD) [20,21]. The smoothing involves a decrease in the number of maxima, initially located on the atoms, and eventually leads to a single maximum located on the SO_2 group of the 1,2-thiazine dioxide ring of the molecule.

8.2.2.2 Application to Electrostatic Potential Functions

The electrostatic potential function $\Phi_M(\mathbf{r})$ generated by a molecule M can be approximated by a summation over its atomic contributions using the Coulomb equation:

$$\Phi_M(\mathbf{r}) = \sum_{i \in M}^{No.\,atoms} \frac{q_i}{|\mathbf{r} - \mathbf{R}_i|} \tag{7}$$

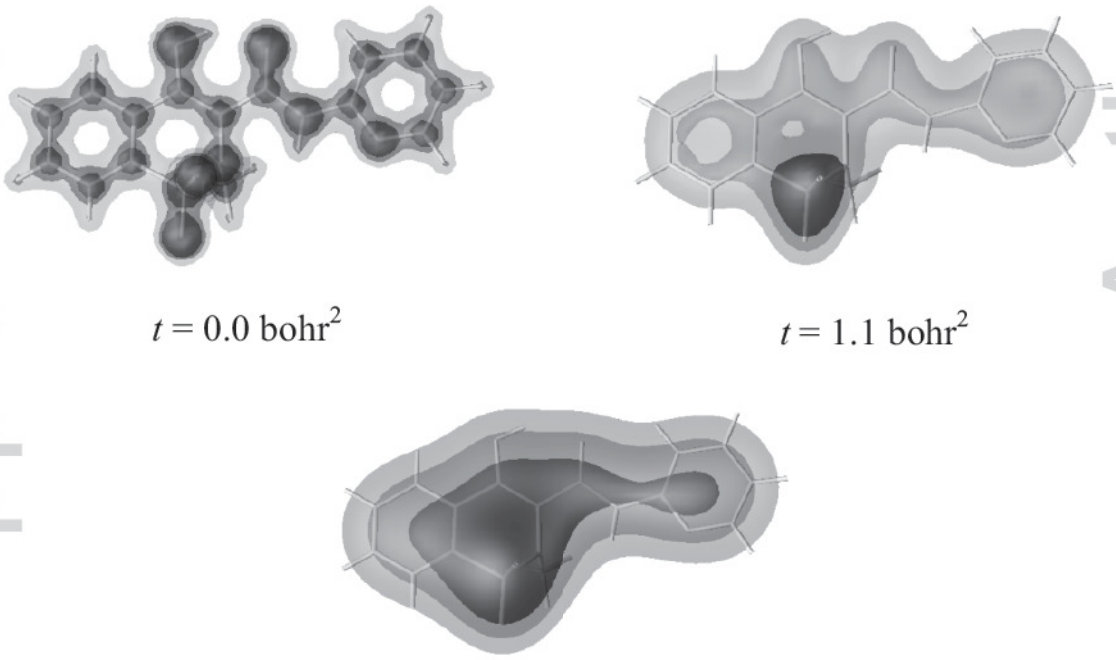


FIGURE 8.3 Iso-contours of the PASA ED distributions of Piroxicam (CSD code: BIYSEH05) calculated at $t = 0.0$ bohr^2 (iso = 0.1, 0.2, 0.3 e$^-$/bohr^3), $t = 1.1$ bohr^2 (iso = 0.1, 0.15, 0.2 e$^-$/bohr^3), and $t = 2.5$ bohr^2 (iso = 0.05, 0.075, 0.10 e$^-$/bohr^3).

where q_i being the net charge of atom i. A smoothed version of the potential generated by atom i, $\Phi_{i,t}(\mathbf{r} - \mathbf{R}_i)$ can be written as:[22]

$$\Phi_{i,t}(\mathbf{r} - \mathbf{R}_i) = \frac{q_i}{|\mathbf{r} - \mathbf{R}_i|} erf\left(\frac{|\mathbf{r} - \mathbf{R}_i|}{2\sqrt{t}}\right) \tag{8}$$

where *erf* stands for the error function. An analytical expression for the corresponding CD function $\rho_{i,t}(\mathbf{r} - \mathbf{R}_i)$ can be obtained from the Poisson equation:

$$-\nabla^2 \Phi_{i,t}(\mathbf{r} - \mathbf{R}_i) = \rho_{i,t}(\mathbf{r} - \mathbf{R}_i) \tag{9}$$

and is expressed as:

$$\rho_{i,t}(\mathbf{r} - \mathbf{R}_i) = \frac{q_i}{(4\pi t)^{3/2}} e^{-|\mathbf{r}-\mathbf{R}_i|^2/4t} \tag{10}$$

In such a formalism, $\rho_{i,t}(\mathbf{r} - \mathbf{R}_i)$ cannot be calculated at $t = 0$. Indeed, that situation corresponds to the original Coulomb potential for which the solution of the Poisson equation is zero. Figure 8.4 shows the evolution of the CD distribution of Piroxicam as t increases from 0.05 to 2.0 bohr^2. Atomic charges q_i were calculated using the restrained electrostatic potential (RESP) method applied at the Hartree-Fock (HF) 6–31G* level and calculated with the program Gaussian09 [23]. From a situation where the maxima and minima of the electrostatic potential are located at the level of the atoms, one evolves toward a decrease in the number of extrema which can be located away from the molecular structure.

8.2.3 CRYSTALLOGRAPHY-BASED METHOD APPLIED TO THE ELECTRON DENSITY

Within the crystallographic approach, an ED distribution function $\rho(\mathbf{r})$ is written as the Fourier transform of the structure factors $F(\mathbf{h})$:

$$\rho(\mathbf{r}) = \frac{1}{V} \sum_{\{\mathbf{h}\}} F(\mathbf{h}) e^{-2\pi i \mathbf{h}.\mathbf{r}} \tag{11}$$

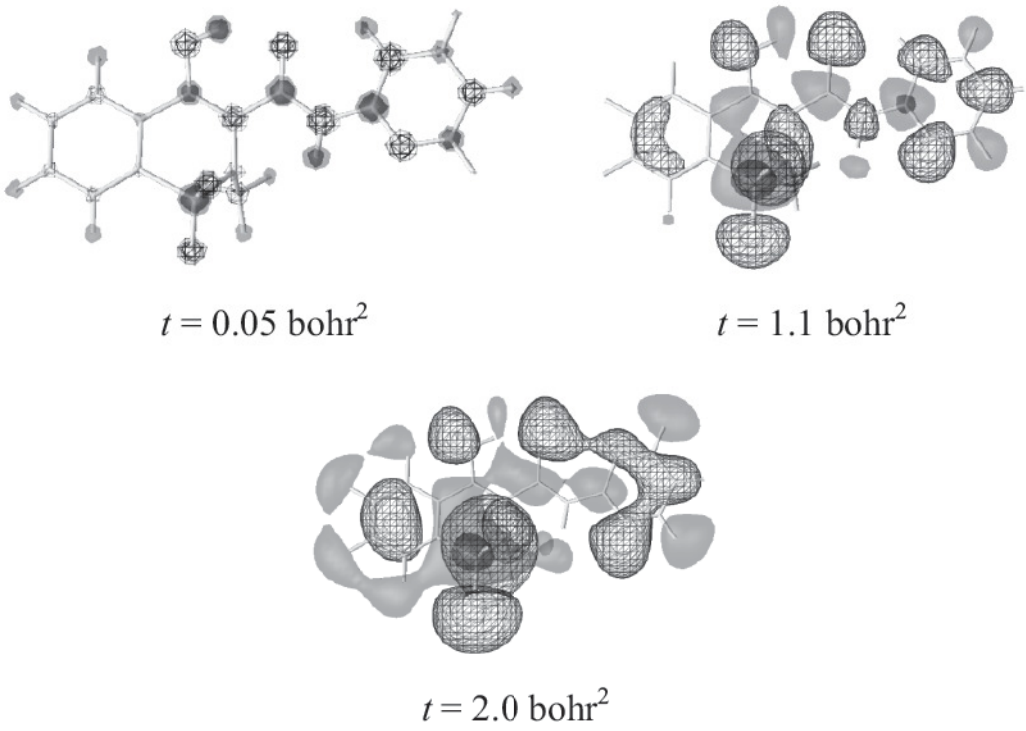


FIGURE 8.4 Iso-contours of the CD distributions of Piroxicam (CSD code: BIYSEH05) calculated using RESP charge values q_a obtained using the program Gaussian09 [22] (HF 6–31G* level) at $t = 0.05$ bohr^2 (iso = –0.25, –0.10, 0.10, 0.25 e$^-$/bohr^3), $t = 1.1$ bohr^2 (iso = –0.0025, 0.0025, 0.01 e$^-$/bohr^3), and $t = 2.0$ bohr^2 (iso = –0.0005, 0.0005, 0.003 e$^-$/bohr^3). Negative and positive iso-contours are displayed using meshes and plain surfaces, respectively.

where V is the volume of the unit cell and **h** is a reciprocal space vector. The structure factors $F(\mathbf{h})$ are mathematically expressed as:

$$F(\mathbf{h}) = \sum_{i=1}^{No.\,atoms} f_i e^{-B_i\left(\frac{\sin\theta}{\lambda}\right)^2} e^{2\pi i \mathbf{h}.\mathbf{R}_i} \tag{12}$$

where f_i is the atomic form factor of atom i, and B_i is the corresponding isotropic temperature factor. Such ED maps can be calculated at various resolution levels using crystallography programs such as XTAL [24]. In practice, the number of known structure factors occurring in Eq. (11) is not infinite and varies with the resolution.

In crystallography, the resolution d_{min} is a well-known concept which is defined using Bragg's law:

$$\left(\frac{\sin\theta}{\lambda}\right)_{\max} = \frac{1}{2d_{\min}} \tag{13}$$

where 2θ is the angle between the diffracted and the primary beams of wavelength l, and d_{min} depends on different parameters including the quality of the crystal, the chemical composition, the radiation used, and the temperature of the experiment. Figure 8.5 depicts the crystallographic ED distribution of the Piroxicam molecule calculated from tabulated f_j factors for independent atoms using the program XTAL [24] at two resolution levels.

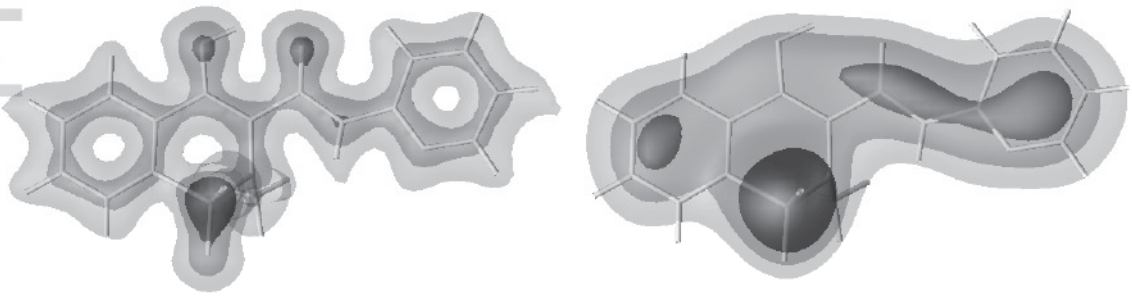
If one considers that the so-called overall isotropic temperature factor B is equivalent to $8p^2u^2$, where u^2 is the mean square atomic displacement, it is found that $u^2 = 2t$ [25]. The crystallographic resolution d_{min} thus differs from the smoothing parameter t which is here related to the dynamical quantity u^2.

8.2.4 WAVELET-BASED MULTI-RESOLUTION ANALYSIS

Wavelet-based techniques do not require the treated signal to be a Gaussian function as needed with the blurring method described above. In practice, wavelet theory is commonly applied to the treatment of signals and images, e.g., in analytical chemistry [26–29], or in spectroscopy [30, 31], but applications to bioinformatics and computational biology [32] as well as to chemistry [33, 34] and chemometrics [34–36] have been reported.

8.2.4.1 Wavelet Transforms

A wavelet transform (WT) is a localized transform in both space (or time) and frequency which uses integration kernels called wavelets. A basis set of wavelet functions $\{Y_{ab}(x)\}$ is built on translated and dilated versions of a so-called mother wavelet $Y(x)$:



1.5 Å 3.0 Å

FIGURE 8.5 Iso-contours of the crystallography-based promolecular ED distributions of Piroxicam (CSD code: BIYSEH05) calculated using the program XTAL [24] at a resolution of (left) 1.5 Å (iso = 1.0, 3.0, 5.0 $e^-/Å^3$) and (right) 3.0 Å (iso = 1.0, 1.5, 2.0 $e^-/Å^3$).

$$\Psi_{ab}(x) = \frac{1}{\sqrt{a}} \Psi\left(\frac{x-b}{a} \right) \text{ with } a \in R_0,\ b \in R \tag{14}$$

where a is the scaling parameter which allows to capture changes in frequency, and b is the shift along the x axis applied to analyze space (time)-dependent variations of a signal. The projection $<f,\Psi_{ab}>$ of a square integrable signal $f(x)$ onto this basis according to:

$$\left\langle f, \Psi_{ab} \right\rangle = \int_{-\infty}^{+\infty} f(x) \Psi_{ab}^{*}(x)\, dx \tag{15}$$

is the result from a continuous wavelet transform (CWT). Y is often required to have a certain number p of vanishing moments:

$$\int_{-\infty}^{+\infty} x^n\ \Psi(x)\, dx = 0 \text{ with } n = 0,\ 1,\ \ldots,\ p\text{-}1 \tag{16}$$

where p is also known as the order of Y. In Figure 8.6, one illustrates the absolute values of the CWT coefficients obtained from the analysis of a-helix propensity descriptors [38] of the amino acids that constitute three ubiquitin ligands (Figure 8.1). The descriptors are reported under the name BLAM930101 in the amino acid database AA index that is publically available [39]. BLAM930101 is found to be at the center of a cluster of a-propensity descriptors of AA index [40]. The initial signal is progressively smoothed up to the point where a limited number of minima are obtained. For Vps27 UIM-1, a single minimum appears at scale a = 16, around Ile267 (residue 13), and is further stabilized at Glu268 (residue 14) at larger scales. For iota UBM1, four minima occur at scale a = 13, at the level of Leu63, Asp71, Asp81, and Lys94 (residues 2, 10, 20, and 33). For the Rabex-5 fragment, Cys19-Lys20, Gly27, Cys38-Trp39, Gln50, and Gln59-Glu67 (residues 5–6, 14, 25–26, 37, and 46–54), correspond to minima in the signal transformed at scale a = 13. It is a scale value that corresponds to a pronounced shift of the coefficient minima along the residue axis. These specific locations will be discussed in the Section 8.3.4.2.

For numerical purposes, the CWT can be discretized, by restricting the parameters a and b to the points of a dyadic lattice. Thus, if a and b/a are set equal to 2^{-j} and 1, respectively, Eqs. (14) and (15) are written as:

$$\Psi_{jk} = 2^{j/2} \Psi(2^j - k) \text{ with } j,k \in Z \tag{17}$$

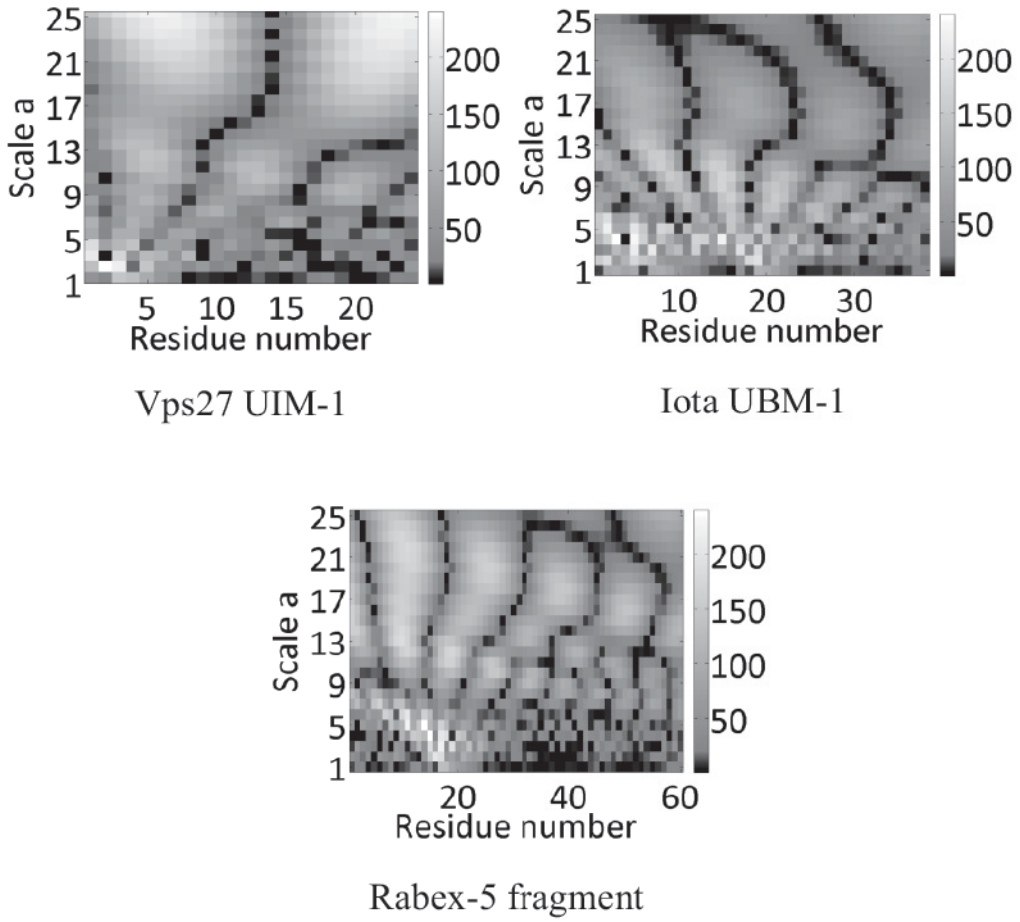


FIGURE 8.6 Absolute value of the wavelet coefficients obtained from the CWT using Y = D15 applied to α-helix propensity descriptors of three Ubiquitin ligands. (Top left) Vps27 UIM-1 (PDB code: 1Q0W), (top right) iota UBM1 (PDB code: 2MBB), and (bottom) a bovine Rabex-5 fragment (PDB code: 2FID). Calculations were achieved using MATLAB [37].

$$f(x) = \sum_{j,k} \left\langle f, \Psi_{jk} \right\rangle \Psi_{jk}(x) \tag{18}$$

The discrete wavelet transform (DWT) of $f(x)$ is calculated by passing the signal through two filters, i.e., a low-pass filter F to obtain the convolution of $f(x)$ with F, and a high-pass filter Y to generate the details, or wavelet coefficients. The procedure can be iteratively repeated by applying the decomposition to the first convolution product, and so on.

8.2.4.2 Multi-Resolution Analysis

A wavelet multi-resolution analysis (WMRA) is a mathematical construction used to express an arbitrary function $f \square L^2(R)$ at various levels of detail.

The function $f(x)$ is developed as in Eq. (18) where $\langle f, \Upsilon_{jk} \rangle$, also written d_{jk}, are called the wavelet coefficients. In practice, the wavelet expansion is truncated at a scale J:

$$f(x) = \sum_k c_{Jk} \Phi_{Jk}(x) + \sum_{j=J}^{J_0-1} \sum_k d_{jk} \Psi_{jk}(x) \text{ with } c_{Jk} = \langle f, \Phi_{Jk} \rangle \tag{19}$$

where it is chosen here to set the resolution of the original signal J_0 equal to zero. Coefficients c_{Jk} are projections of the function f onto a space built on the basis set $\{\Phi_{Jk}(x)\}$. Thus, lower resolution signals are characterized by negative values for J. In Eq. (19), the first sum is a coarse representation of f, where f is replaced by a linear combination of a finite number of translations of the scaling functions $\Phi_{J0}(x)$. The remaining terms are refinements (details) determined at each scale j by translations of the wavelet $\Psi_{j0}(x)$ that are added to obtain a successively more detailed approximation of $f(x)$. For example, B-spline functions can be considered as scaling functions as illustrated by Stollnitz et al. [41,42] and applied by Carson [5] to model protein backbones and DNA surfaces at various levels of resolution.

A fast and accurate algorithm due to Mallat, named the 'pyramid algorithm' or the fast wavelet transform (FWT) [43], is applicable to signals consisting of 2^n data points. Its aim is to derive a mapping between the sequence $\{c_j\}$ and the sequences $\{c_{j-1}\}$ and $\{d_{j-1}\}$ through the following identities [44,45]:

$$c_{j-1,l} = \sum_k h_k c_{j,2l+k} \tag{20}$$

$$d_{j-1,l} = \sum_k g_k d_{j,2l+k} \tag{21}$$

where the numbers h_k are called the filter coefficients, and the wavelet coefficients g_k are obtained directly from the filter coefficients h_k [46]:

$$g_k = (-1)^k h_{k_{\max}-k}, \text{ where } k = 0, 1, \ldots, k_{max} \tag{22}$$

Equations (20) and (21) are further applied to the sequence $\{c_{j-1}\}$ in order to obtain the new sequences $\{c_{j-2}\}$ and $\{d_{j-2}\}$. This procedure is repeated until the full FWT is achieved. The full procedure is named the cascade

algorithm. Equation (21) shows that the calculation of coefficients $\{c_{j-1}\}$ from coefficients $\{c_j\}$ implies a downsampling, i.e., the number of coefficients is reduced by 2. It is also known as a decimated wavelet analysis. On the contrary, reconstruction implies an upsampling procedure, i.e., the number of data point is multiplied by 2 at each level of resolution.

The inverse mapping can be derived according to:

$$c_{jl} = \sum_k h_{l-2k} c_{j-1,k} + \sum_k g_{l-2k} d_{j-1,k} \tag{23}$$

The inverse FWT is obtained by repeated application of this equation for $j = J+1, J+2, \ldots$, up to J_0.

For example, Main and Wilson used an inverse WT approach to increase the resolution of ED maps [47]. The method proposed by the authors is based on a preliminary design of histograms of wavelet coefficients obtained from one-level WT decomposition of identified ED maps. Then, their procedure consists, as briefly summarized, of the following steps: (i) a one-level WT decomposition of a low-resolution ED map, (ii) the creation of an ordered list of the wavelet coefficients, (iii) a match of the wavelet coefficients with the preliminary obtained histograms, (iv) an inverse WT to generate the higher resolution ED map.

8.2.4.3 Multidimensional Cases and Smoothing

A simple way to obtain wavelet coefficients in dimensions higher than one is to carry out a 1D wavelet decomposition for each variable separately. The standard decomposition, described by Stollnitz et al., consists in the application of a 1D FWT to each row of data values [41,42]. The operation gives, for each of them, an average signal along with detail coefficients. Next, these transformed rows are treated as if they formed an image, and a 1D FWT is applied to each column. In the non-standard decomposition scheme, operations in rows and columns alternate. One applies a one-level decomposition to each row, followed by a one-level decomposition to each column. Then, one repeats the process on the resulting filtered image, and so on.

To obtain an image at various levels of decomposition, a smoothing procedure is required, which is applied before reconstructing the original signal as follows: all details generated after a given number of decomposition levels using a FWT are set equal to zero before a full reconstruction procedure

is applied to restore the original number of data points. An example applied to the 3D promolecular ED of Piroxicam using the Daubechies' wavelet of order 10, D10, is displayed in Figure 8.7. Similarly to the crystallography-based approach, smoothing the PASA ED grid involves a decrease in the number of maxima which are initially located on the atoms, then on the rings and heteroatoms. The procedure eventually leads to a single maximum.

In the so-called 'à trous' algorithm, the smoothing procedure is implemented as a convolution product with a symmetric filter. The corresponding WT is achieved by inserting zeroes between the h_k coefficients [48]. The algorithm is illustrated by Gonzalèz-Audicana et al. in a paper comparing the Mallat and the 'à trous' algorithms [49], and relations between the two approaches are discussed by Shensa [50]. While the Mallat algorithm leads, at each resolution level, to a decrease in the number of points in an image, that number remains constant with the 'à trous' algorithm. As already mentioned for the Mallat algorithm, a reconstruction is necessary to preserve the number of data points in an image. The application of the 'à trous' algorithm to the promolecular ED grid of Piroxicam using the Daubechies' wavelet of order 10, D10, is illustrated in Figure 8.8. As in Figure 8.7, at the wavelet transformation level $J = -5$, there is only one maximum left, which is slightly displaced away from the SO_2 group. That single maximum was located on the sulfur atom with the Gaussian blurring and crystallography-based approaches, which thus appear to be more sensitive to the atomic number of the heaviest atom.

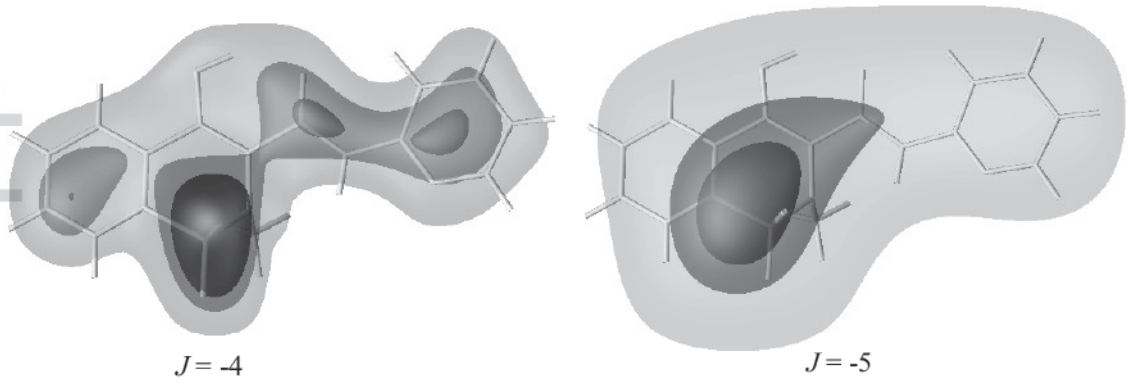


FIGURE 8.7 Iso-contours of the smoothed PASA ED distributions of Piroxicam (CSD code: BIYSEH05) calculated using the FWT approach with Φ = D10. (Left) $J = -4$ (iso = 0.1, 0.2, 0.25 $e^-/Å^3$) and (right) $J = -5$ (iso = 0.03, 0.1, 0.125 $e^-/Å^3$). The grid size is 2^8 x 2^8 x 2^8 and the grid interval is 0.125 Å.

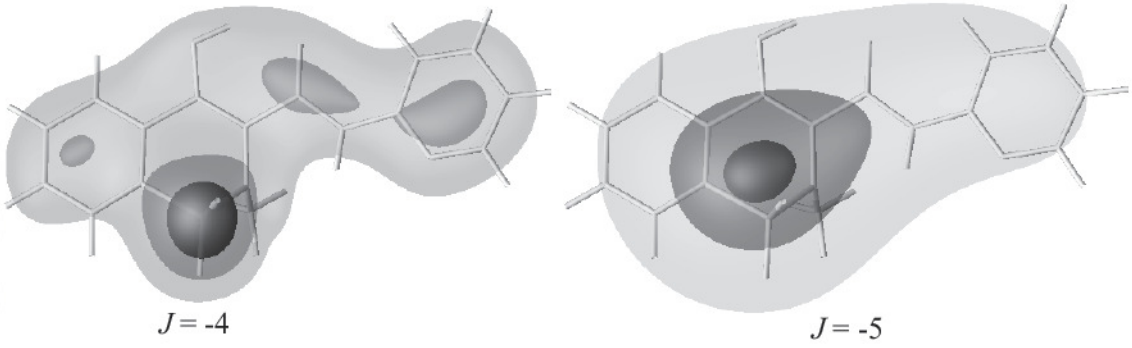


FIGURE 8.8 Iso-contours of the smoothed PASA ED distributions of Piroxinam (CSD code: BIYSEH05) calculated using the 'à trous' approach with the Lagrangian-based D10 filter. (Left) $J = -4$ (iso = 0.1, 0.2, 0.3 $e^-/Å^3$) and (right) $J = -5$ (iso = 0.05, 0.1, 0.125 $e^-/Å^3$). The grid size is 2^8 x 2^8 x 2^8 and the grid interval is 0.125 Å.

8.2.5 MOLECULAR DYNAMICS-RELATED APPROACHES

In addition to direct smoothing procedures described, e.g., by Eqs. (2) and (19), specific approaches are also currently applied to artificially smooth functions such as complex potential energy hyper-surfaces (PES), to allow a molecular system to visit less probable energy wells in a faster way during a MD simulation. Rather than modifying the force field (FF) itself, like it is done in potential smoothing approaches, the FF is biased by an extra term. This can be done either by reducing the energy barriers, through the hyper-dynamics approach, or by progressively filling the already visited energy wells, through the metadynamics and flooding approaches.

8.2.5.1 Hyperdynamics

The aim of hyperdynamics is to build an auxiliary system, which actually is the original system with a faster dynamics [51–53]. A bias potential $\Delta V(\mathbf{r})$ is added to the original potential energy function $V(\mathbf{r})$, which allows a reduction of the height of the energy barriers:

$$\Delta V(\mathbf{r}) = \begin{cases} 0 & \text{if } V(\mathbf{r}) \geq E \\ \dfrac{(E - V(\mathbf{r}))^2}{\alpha + E - V(\mathbf{r})} & \text{if } V(\mathbf{r}) < E \end{cases} \tag{24}$$

The approach thus requires a bias potential characterized by E and a that control the depth and flatness of the biased potential wells,

respectively. At each time step of the simulation, the system undergoes forces that correspond to an increased potential energy value, allowing it to more easily cross energy barriers. This is illustrated in Figure 8.9 for a hypothetical 1D potential energy curve. The figure shows the effect of decreasing the value α on the energy barrier occurring between two potential wells. With respect to the Gaussian blurring method (Figure 8.2), the two potential wells stay preserved but are leveled out in the hyperdynamics approach.

8.2.5.2 Metadynamics

As for hyperdynamics, an auxiliary system is built as being the original system driven by collective variables $s(\mathbf{r})$. Such variables can be distances, torsion angles, a coordination number, lattice vectors, a solvation energy, a root mean square deviation (RMSD), etc. depending upon the system under consideration. A time-dependent bias potential $\Delta V(s(\mathbf{r}),time)$ is added to the original potential function $V(\mathbf{r})$ to avoid the system visiting already explored regions in the collective variable space [54, 55]. Doing so, at each time step, the potential energy involves a sum of extra Gaussian-type contributions, each depending on a previously visited state:

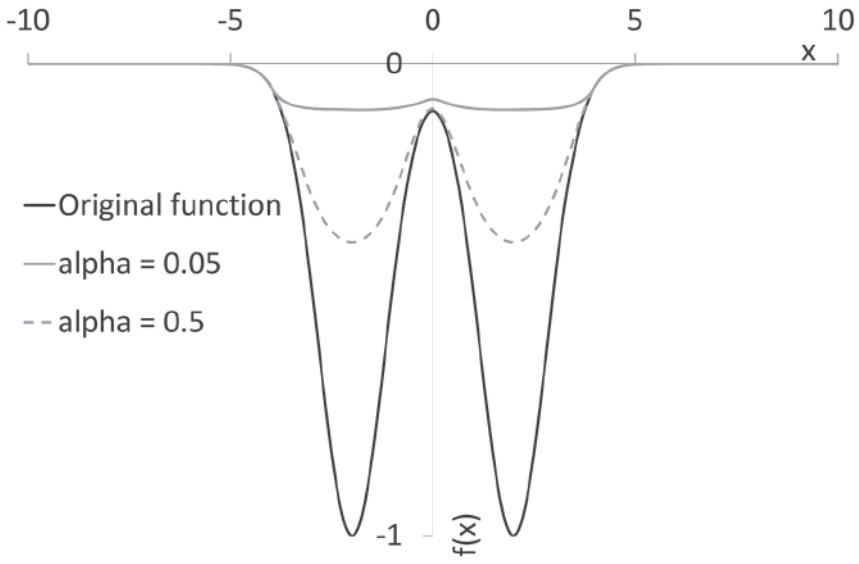


FIGURE 8.9 Effect of the application of a hyperdynamics bias to a hypothetical 1D potential energy function $V(x) = -(e^{-0.75*(x-2)2} + e^{-0.75*(x+2)2})$ with $E = -0.05$ (arbitrary units).

$$\Delta V_i(s(\mathbf{r}),time) = \sum_{i=1}^{\text{all previous MD steps}} w\exp\left(-\frac{\|s(\mathbf{r})-s(\mathbf{r}_i)\|^2}{2\delta s^2}\right) \quad (25)$$

This can be visualized as a well filling procedure, as nicely illustrated in Figure 1 of the paper by Barducci et al. [55].

To limit the total number of PES minima, or to avoid having to visit several times the minima as done, for example, when using Monte Carlo (MC) or MD approaches, the principle of flooding energy minima can be applied until the system finds a way towards the global minima. When carrying out MD simulations, the kinetic energy can be adjusted [56]. In a MC procedure, the so-called basin hopping method was proposed which consists in the transformation of the PES such that all the potential energy values characterizing an energy well are replaced by the single minimal energy value [57]. The PES thus adopts a staircase shape where local energy barriers are neglected.

8.2.6 DISCRETIZATION TECHNIQUES

In this section, procedures used to replace a discrete or continuous molecular property by a limited number of discrete data points are briefly presented.

The so-called vector quantization (VQ) algorithms are well-known in data compression processes. They approximate an initial distribution probability of data points by a set of representative vectors. The methodology consists in the partitioning of the initial space into compact and well-separated clusters of points in such a way that all data points in one group are replaced by a single representative data point for that particular group. Each representative point is named a code vector. Clusters can, for example, be estimated from distance criteria, by a tessellation technique like Delaunay triangulation, … or by convolution with a Gaussian function [58]. De-Alarcón et al. applied their technique to the lowering of resolution of van der Waals surfaces of proteins and to cryo-EM maps [58]. Starting from an initial low-resolution map, they obtain a limited number of pseudo-atoms (code vectors), each characterized by a probability distribution function. Volumetric elements, named alpha shapes, are associated with the pseudo-atoms of the map and are used for the detection of deep clefts and channels in the protein system. Wriggers et al. [59, 60] proposed VQ procedures, implemented in the package Situs, to quantize two EM grids, at an atomic and a lower resolution, to rapidly enable the search for the

best match between the two so-obtained VQ representations. In partial relation with VQ, Vorobjev proposed a method to locate low-resolution binding sites of a protein from its solvent accessible surface (SAS) represented by a set of discrete dots [61]. Binding sites result from the clustering of SAS-related dots followed by the location of centers of dense clusters.

In their approach, Glick et al. [62, 63] represented small molecules through a limited number of points obtained using a clustering of atoms based on their separating distances. The authors developed a method for ligand binding site identification on a protein. A hierarchy of models generated using a k-mean clustering algorithm for the ligand under consideration is established starting from the lowest resolution representation of the ligand, i.e., one single point located at the mean position of the ligand atoms. The resulting graphs were used in ligand-docking applications and illustrated the decrease in the possible number of conformers.

An inverse procedure, starting from all the atoms of a molecule, was implemented by Leherte et al. to locate critical points (CP), i.e., points where the gradient of the 3D field vanishes, in smoothed molecular fields [64]. It is based on the work of Leung et al. whose algorithm was originally established to cluster data by modeling the blurring effect of lateral retinal interconnections based on scale space theory [65]. The various steps of the resulting algorithm are as follow. (i) At scale $t = 0$, each atom of a molecular structure is considered as a local extremum of the molecular field to be analyzed. All atoms are then considered as the starting points of trajectories whose merging procedure is described hereafter. (ii) As t increases from 0.0 to a given maximum value, each extremum moves continuously along a path to reach a location in the 3D space where the gradient of the molecular field is zero. As t increases, trajectories progressively merge to CP locations. It is thus possible to assign, to each CP, a number of atoms which correspond to the starting points of the merged trajectories. (iii) The procedure can be carried out until the whole set of extrema becomes one single point. This is the ultimate stopping criteria of the merging procedure. Various applications can be found in previous publications [25, 66–69].

To analyze low-resolution ED grids obtained using crystallography-based approaches, we used the program ORCRIT, that was developed by Johnson [70]. The information generated by the topological analysis method implemented in ORCRIT allowed, for example, as sign ellipsoids centered at the CPs of ED grids of biomolecular systems to probe the interaction potential between a ligand and a DNA fragment [71], and between

protein-DNA partners [72, 73], or to generate descriptors for small molecules [74, 75].

8.3 APPLICATION FIELDS

In this section, wherein some application domains are presented, a distinction is made between the smoothing of continuous or pseudo-continuous functions as in global optimization, denoising, and molecular similarity applications, and discrete representations, for example, in coarse-graining studies. In that latter sub-section, specific applications of RPCMs to MD simulations of proteins of ubiquitin complexes are also reported.

8.3.1 GLOBAL GEOMETRY OPTIMIZATION

Global optimization is one of the major fields of research that may require the use of smoothed or low-resolution molecular properties [76]. Its aim is to optimize a function considering some constraints, e.g., finding the global minimum of a potential energy hypersurface. Constraining degrees of freedom and/or reducing the level of detail are helpful ways to tackle multiple minima problems. Since they facilitate the overcome of energy barriers by reducing the number and depth of energy wells, global geometry optimization techniques are widely used, e.g., to generate atom clusters [56, 77]. A method to smooth interaction potential functions like well-known FFs is based on the results of the DEM [78], as for example implemented in the program package TINKER [79]. It consists in calculating a convolution product of the original energy function with a Gaussian as described earlier in the papers [8, 22, 77, 80–82]. As energy terms of conventional FFs are not always suitable for direct applications of the diffusion equation, they need to be replaced by, e.g., Gaussian approximations [8]. Mathematical formalisms and applications to Ar clusters, small molecules, and docking of α-helices were treated by Pappu et al. [78], and a study regarding a short peptide was proposed by Hart et al. [22]. Other mathematical approaches used to reduce the number of energy minima were reviewed by Schelstrate et al. [80] and different smoothing functions were proposed by Grossfield and Ponder [81] and Shao et al. [83].

In molecular distance geometry problems, which consist in the determination of a molecular structure from a set of interatomic distances,

global optimization techniques can be based on a smoothing of the objective function. Various smoothing procedures are proposed in literature [7, 84–86]. The Gaussian transformation however appears to be mostly used [7, 84, 85].

The case of molecular docking applications, i.e., the search for an optimal arrangement of molecular partners, asks for a scoring function, which is often selected to be the potential energy of the system. For example, in the so-called 'stochastic approximation with smoothing', the mathematical aspects of the implementation differ from the DEM, but the main philosophy is similar in the sense that one initially looks for a single minimum in a smoothed energy hypersurface and then one iteratively recovers the initial resolution of the energy function while performing a minimization calculation at each step [87].

8.3.2 MOLECULAR SIMILARITY

Global optimization algorithms are often sought to tackle molecular similarity problems, which also involve many local solutions. A recent review of similarity-based methods was, for example, proposed by Cai et al. [88]. A way to reduce the number of possible alignments is to lower the level of detail of the molecular field under consideration [89–92]. To compare molecular structures or surfaces, it is useful to model the molecular properties to be compared using mathematical functions that allow fast calculations involving the maximization of a similarity score or the minimization of a distance-based score. The use of Gaussian functions for the evaluation of the molecular similarity has been an attractive strategy as it both allows short calculation times and it is easy to implement [93]. Indeed, using such functions, similarity measures are directly related to distances between the atoms that constitute the molecular structures to be compared [94, 95].

In our works about the superposition of small drug molecules [91, 92], we used well-known similarity indices and applied them to smoothed ED and CD distribution functions. We also evaluated molecular similarity by representing molecular systems as graphs of CPs obtained from smoothed ED distributions [74, 96].

Carbó-Dorca and co-workers were the first to report the now widely used quantum molecular similarity measure (QMSM) Z_{AB} definition between

two chemical systems characterized by ED distribution functions ρ_A and ρ_B [97–99]:

$$Z_{AB} = \int\int \rho_A(\mathbf{r}_1)\, O(\mathbf{r}_1,\mathbf{r}_2)\, \rho_B(\mathbf{r}_2)\, d\mathbf{r}_1 d\mathbf{r}_2 \tag{26}$$

Depending upon the nature of the operator $O(\mathbf{r}_1,\mathbf{r}_2)$, overlap-, Coulomb-like, ... similarities are obtained [94, 100]. The similarity measures can be combined to lead to similarity indices, such as the well-known Cosine-like (also known as Carbó), Tanimoto, or Hodgkin-Richards indices [93, 101–103]:

$$S_{AB,Carbó} = \frac{Z_{AB}}{\sqrt{Z_{AA} Z_{BB}}} \tag{27}$$

$$S_{AB,Tanimoto} = \frac{Z_{AB}}{Z_{AA} + Z_{BB} - Z_{AB}} \tag{28}$$

$$S_{AB,Hodgkin-Richards} = \frac{2Z_{AB}}{Z_{AA} + Z_{BB}} \tag{29}$$

As already mentioned in the Section 8.2, Carbó-Dorca and coworkers developed the so-called atomic shell approximation (ASA) method, in which atomic or molecular ED are expressed as linear combinations of 1s Gaussian-type functions centered at atomic positions [104]. Such approximations were shown to be useful in QMSM calculations, especially to model promolecular ED distribution functions [105] where the EDs are expressed as sums over the contributions of independent atoms [18, 106, 107]. The use of smoothed PASA models obtained using Eq. (5) allows to significantly reduce the number of local solutions as illustrated in Figure 8.10 for the alignment of $C_2(CN)_4$ (CSD code: TCYETY) onto acridine (CSD code: ACRDIN01), generated using $S_{AB,Tanimoto}$ together with the overlap integral:

$$Z_{AB} = \int\int \rho_A(\mathbf{r}_1)\ \rho_B(\mathbf{r}_2)\, d\mathbf{r}_1 d\mathbf{r}_2 \tag{30}$$

For the unsmoothed case, i.e., at t = 0 bohr^2, multiple maxima are obtained. Their number is drastically reduced to four at t = 0.1 bohr^2, and a

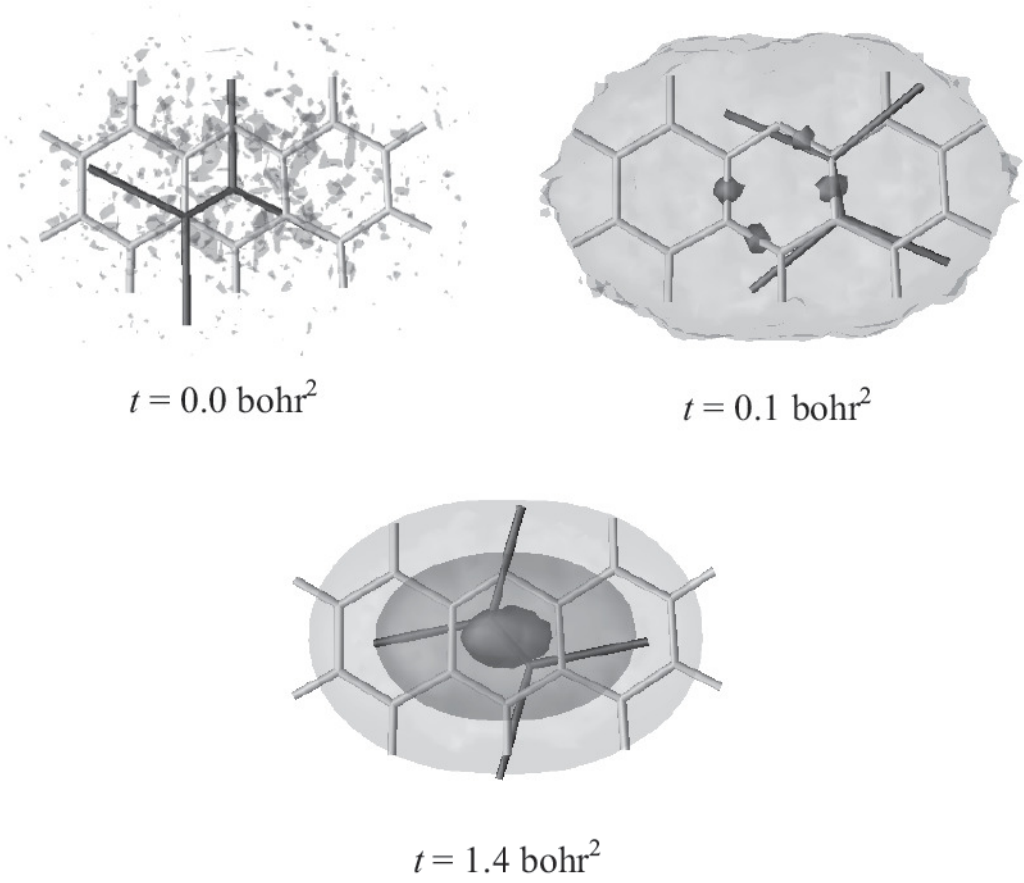


FIGURE 8.10 Iso-similarity contour maps calculated from the alignment of the PASA ED of $C_2(CN)_4$(CSD code: TCYETY) (black sticks) on Acridine (CSD code: ACRDIN01) (light gray sticks) using the overlap integral measure combined with the Tanimoto index, obtained at $t = 0.0$ bohr^2 (iso = 0.04, 0.07), 0.1 bohr^2 (iso = 0.1, 0.3), and 1.4 bohr^2 (iso = 0.3, 0.5, 0.7). To generate the maps, a grid is defined around the largest ligand. The center of mass of $C_2(CN)_4$ is placed at each grid point and its optimal orientation is determined as corresponding to the maximum value of $S_{AB,Tanimoto}$ at that point. The example is taken from the test cases studied in Constans et al. work [108].

single maximum $S_{AB,Tanimoto}$ value of 0.74 is obtained at $t = 1.4$ bohr^2. At $t = 0$, the atoms tends to be superimposed while at larger smoothing values, the global shape of the molecules are aligned.

Molecular surfaces can also be approximated using low-resolution functions, e.g., spherical harmonics [109]. Ritchie et al. suggested that surface representations which contain too high-resolution details may not be particularly convenient to search for regions of similarity or complementarity between two molecules [109]. The authors therefore proposed the use of low-resolution real spherical harmonics to represent and compare macromolecular surface shapes. Rotations of a molecular surface can thus be simulated by rotating only the harmonic expansion coefficients.

Hakkoymaz et al. adapted well-known molecular similarity indices, like the Carbó-Dorca and Hodgkin indices, to wavelet coefficients [110]. They applied the revised similarity indices to a multi-resolution analysis (MRA) decomposition of electrostatic potential grid points. Rather than working with modified similarity indices, Beck and Schindler used the original indices but applied them to a MRA decomposition of a 3D molecular field, like the ED [111]. Martin et al. applied a similar MRA analysis, but used graph descriptions of the MRA molecular images in molecular alignment applications [112].

The search for similarity degrees through wavelet coefficients is a well-known technique to also compare protein sequences, represented either by their 3D Cα coordinates [113], or by their amino acid sequence [114]. Transmembrane proteins are good study cases since their membrane and non-membrane regions are constituted by sequences of contiguous amino acid residues. In Fisher's work, the smoothing of a hydropathy profile to predict the location of helices in transmembrane proteins is achieved by setting to zero wavelet coefficients associated with high frequencies of the hydropathy signals [114]. The sequences are submitted to a wavelet-based filtering algorithm, that provides smoothed profiles whose reduced number of extrema are identified to transmembrane helices, as also achieved using a CWT by Qiu et al. [115] and by Vannuci and Liò [116]. de Trad et al. proposed a method that is based on the MRA decomposition of a protein sequence signal built from the Fourier transform of amino acid properties like the electron-ion interaction potential [117]. A DWT is then applied to the transformed sequences in order to decompose the data into a number of levels. Two protein sequences can be compared at each level through the calculation of a cross-correlation coefficient, which is seen as a similarity score. Sabarish and Thomas applied a similar approach to protein sequence similarity and functional classification [118]. They used a DWT to decompose profiles of amino acid properties. It was followed by a correlation analysis that allows the classification of protein sets into functional classes. The conservation of physico-chemical properties in a functionally similar family of proteins was also studied by MRA [119]. Wen et al. applied 46 kinds of wavelets to a set of protein profiles, at various levels of decomposition, and defined a metrics to quantitatively evaluate the similarity degree between any two protein sequences with low identity [120]. In addition to the calculation of cross-correlation analysis, the comparison of protein profiles reconstructed after zeroing detail coefficients can be

quantified using a distance-based criteria, such as in the work of Krishnan et al. [121].

The MRA technique was also applied to DNA strands, transformed to a sequence of integers (Adenine = 1, Thymine = 2, Cytosine = 3, Guanine = 4), by Tsonis et al. who detected localized periodic patterns and suggested DNA construction rules [122]. Machado et al. used complex numbers (Adenine = $1 + i0$, Thymine = $0 + i1$, Cytosine = $-1 + i0$, Guanine = $0 - i1$) to encode human DNA and showed that the Shannon continuous wavelet led to interpretable patterns [123], while Mena-Chalco et al. decomposed DNA strands into four binary sequences, one for each nucleic base, so as to consider a single descriptor for each base [124].

8.3.3 DENOISING OF SIGNALS AND IMAGES

Smoothing and denoising are related by the fact that the methods aim at separating useful and useless information in a data set or a signal [34]. In a smoothing approach, high-frequency components are removed while, in a denoising approach, low-amplitude components are removed [125].

In a DWT approach, smoothing of a signal is achieved through the following four steps: (i) transform the signal up to a selected level, (ii) recognize the wavelet coefficients associated with the highest frequencies, i.e., the detail coefficients, (iii) cancel those coefficients, and (iv) apply an inverse WT to the resulting signal. It has, for example, been achieved to smooth PASA ED or QM ED distribution functions of drug molecules in order to further generate a limited amount of CPs used in molecular alignments [74].

Denoising is achieved by canceling wavelet detail coefficients that are lower than a threshold value. This is called hard tresholding. One can distinguishes between hard and soft tresholding where values slightly below the threshold are not set to zero but attenuated so as to obtain smoother transitions between the original and the deleted values. These two approaches were tested by Pilard and Epelboin in their work about the restoration of noisy X-ray topographs [126]. Several threshold values and methods are presented by Ergen [127] to denoise heart sounds, and by Jeena et al. [128]. Non-reconstructed signals can also be useful, as shown by Chen et al. in their work about the parametrization of CG potential energy functions of DNA [129]. The authors decompose all-atom distance,

angle, ... probability distributions calculated from all-atom MD simulations to generate the corresponding stretching, bending, and non-bonded coarse-grained versions.

Simulating proteins at low-resolution is a way to overcome structural inaccuracies issued from NMR data or from an approximate model. In their paper, Vakser et al. digitized a protein image onto a 3D grid [130]. Any grid point outside the molecular volume was set equal to 0, otherwise it was set equal to a numerical value corresponding to the protein surface or to the protein core, as also applied by Katchalski-Katzir et al. in a docking procedure [131]. The structural elements smaller than the grid interval are thus eliminated from the initial protein structure. The approach is implemented in the program GRAMM (Global RAnge Molecular Matching) reviewed together with other docking techniques by Russell et al. [132]. Vakser and colleagues studied large sets of protein complexes, at various resolution levels, and showed that low-resolution docking can provide gross structural features of protein-protein organization [133, 134].

Denoising procedures are also useful to locate water molecules in experimental ED maps. For example, Nittinger et al. [135] first generate a Gaussian expression for the ED associated with water molecules from a set of PDB structures and analyze it to classify the water molecules. They suggest that the procedure could be extended to detect misleadingly placed water molecules through a difference of Gaussian filters characterized each by a different width.

8.3.4 DISCRETE REPRESENTATIONS OF BIOMOLECULAR STRUCTURES AND PROPERTIES

8.3.4.1 Coarse-Grained Representations

For several years, much effort has been put into accelerating computational techniques such as MD and normal mode analysis for simulating large biological systems [136–138]. Enhancements to these well-known algorithmic procedures are based, notably, on a spatial coarse-graining of the molecular structures [139, 140]. Techniques that are relevant to coarse-graining of molecular structures are not necessarily linked to the smoothing of molecular properties, but they are nevertheless based on a decrease in the number of degrees of freedom

and in the level of details. Rather than simulating the molecules at their atomic level, one reduces their description to a limited set of points, either centered on selected sites/atoms such the Cα atoms of a protein backbone [136, 141, 142], on the center of mass of specific groups of atoms like amino acid residues [143], or on a set of merged atoms [144]. A shape-based coarse-graining approach [145, 146], now implemented in the program VMD [3], was proposed to generate highly coarse-grained descriptions of biomolecular complexes like viral capsids, proteins, and membranes. The authors used a reduced set of point masses, determined from a Voronoï-based partitioning of the molecules into domains of atoms, to reproduce the overall shape of the systems while respecting the mass distribution.[145,146] Reviews on the progresses on coarse grain (CG) models can be found in several other references [147–154].

The development of CG interaction potential functions is generally made either from atomistic interaction potential [155] or MD results [156–159], via experimental data such as B-factors [160], or through the fitting of a potential function achieved by matching CG and atomistic distributions [156]. For example, Lyman et al. presented a new method for fitting spring constants to mean square CG-CG distance fluctuations computed from atomistic MD [161]. Orellana et al. also developed an approach to design robust and transferable elastic network models that fit MD simulation data so as to best approximate local and global protein flexibility [162]. CG interaction potential can also be designed by fitting energy functions to smoothed all-atom energy profiles [129], or by applying the Inverse Monte Carlo approach [163], used for iteratively adjusting a CG potential function until it matches a target radial distribution function. Another example is the parametrization of the MARTINI FF, designed to reproduce partitioning free energies between polar and a polar phases of a large number of chemical systems [164, 165]. The model is based on a four-to-one mapping, i.e., four heavy atoms are represented by a single interaction center, except for small ring-like fragments. In the UNRES model, a peptidic chain is represented by a sequence of backbone beads located at peptide bonds, while side chains are modeled as single beads attached to the Cα atoms, which are considered only to define the molecular geometry [166]. In the so-called SimFold CG description and energy function, a mixed representation is used. Residues of aqueous proteins are represented by backbone atoms N, Cα, C, O, and H, and one side chain centroid [167, 168]. The more recent SIRAH force field, applicable to DNA and proteins, is intended to capture temperature and solvent effects without the need of any structural constraints as required in MARTINI. [142]. Multiscale

methods, that combine several levels of description, are also appealing since they allow to model limited regions of space with details while representing the outer regions by coarser models [149, 169, 170]. Multiscaling is not only used at the level of molecular structure representation, but is also applied to solve the atom equations of motion in MD simulations. Particularly, He et al. described the so-called smoothed molecular dynamics (SMD) method wherein the equations of motion are solved for nodes of a grid to generate nodal velocities and accelerations [171]. Atomic velocities and positions are then updated using the properties of the node they are associated with. In such a way, the SMD time step can be much larger than the conventional MD time step, depending upon the grid element size.

The advantage of using non-atomic representations is not limited to the increase of the speed of computation. Simplified representations of protein geometry have also been used by many groups to reduce sensitivity to small perturbations in conformation, e.g., when docking a ligand versus a receptor [172, 173]. Sternberg et al. replaced amino acid residues with spheres of varying size and performed docking to maximize the buried surface area [173].

In DNA, structural elements such as chains and bases can be modeled as rigid segments connected through energetic terms [174–176]. Particularly, von Kitzing and coworkers described an approach which reduces the number of degrees of freedom by assembling certain groups of atoms into configurational structures with less degrees of freedom [174, 175]. Corresponding potential energy functions were constructed with respect to these new variables using methods from the theory of wavelets, splines, and radial basis functions. DNA is, under such a description, represented by two kinds of rigid substructures: the bases and the phosphate groups, and one rotational group: the ribose subunit. In a recent paper, Naômé et al. developed a CG model and interaction potential to accurately reproduce the structural features of the underlying atomistic DNA system [159]. The authors used a one-site representation of the DNA nucleotides together with a limited number of intramolecular and intermolecular pair interaction potentials.

To model small molecules, the grouping of atoms to form pseudo-atoms is often achieved to accelerate the search of substructures in large databases [177, 178], in molecular similarity applications [74], and in docking calculations [62, 63]. Among the most recent review papers in the field of reduced graph representation of small molecules, one can cite, for example, the work of Birchall and Gillet [179].

8.3.4.2 Reduced Point Charge Models of Proteins

In previous works, we described an approach to model, through MD simulations, protein systems using an hybrid Amber99SB FF mixing all-atom bonded and van der Waals terms with reduced point charge sets [67–69]. Such a model allows to preserve the information regarding the atomic positions and does not require any back-mapping procedure to recover the all-atom structure of the system. It also eliminates some drawback related to the use of CG models, such as structure collapsing [180].

The RPCMs involve, for each amino acid, a limited amount of point charges located at the extrema of their smoothed positive and negative CD distribution function (Figure 8.11). Cao and Voth also used a CG generation procedure wherein they treat separately positive and negative charges to represent effective dipole moments [181]. Figure 8.11 illustrates positive and negative iso-contours of the CD of the structure Gly-Hisδ-Gly built from Amber99SB atomic charges according to Eq. (10). The CD is smoothed at a level of t = 1.7 bohr^2 and is characterized by two extrema located on the amino acid backbones and three extrema located on the Hisδ side chain.

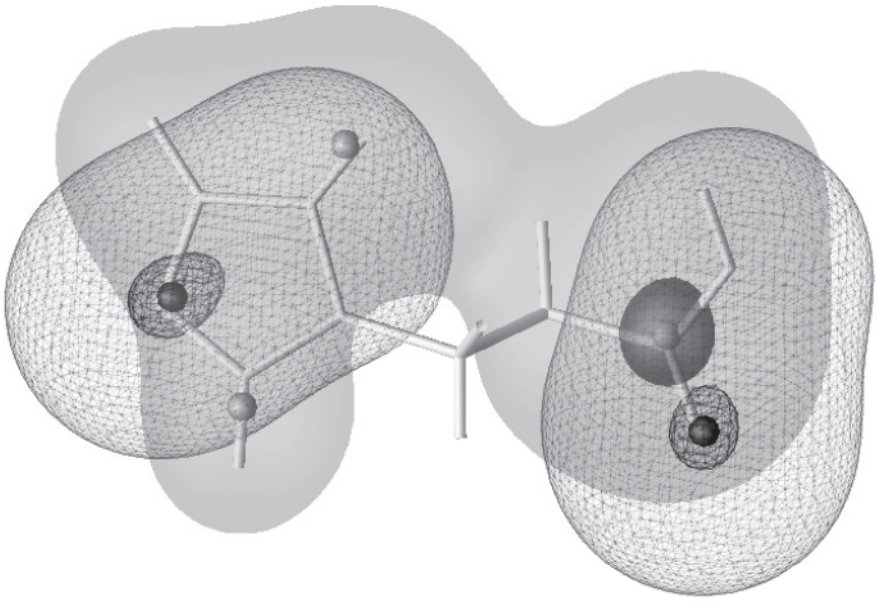


FIGURE 8.11 Iso-contours of the smoothed CD distributions (t = 1.7 bohr^2) built from the positive (plain surface; iso = 0.001, 0.0055) and the negative Amber99SB charges (mesh; iso = –0.001, –0.0055 e$^-$/bohr) for Gly-Hisδ-Gly. Point charges are located at the extrema of the negative (black spheres) and positive CD (gray spheres).

In our last study [69], it was shown that the RPCMs built with charges values fitted to electrostatic forces, and mostly located on selected atoms, led to MD trajectories that are, to some extent, similar to the all-atom ones. Here, we perform MD simulations of three ubiquitin complexes over longer time scales than previously reported to verify the stability of the complexes generated under RPCM conditions. Applications are given for the three ubiquitin complexes Vps27 UIM-1–ubiquitin [182], Iota UBM1–ubiquitin [183], and a bovine Rabex-5 fragment complexed with ubiquitin [184]. Representations of the PDB structures are shown in Figure 8.1. The three ligands interact with a hydrophobic patch of ubiquitin centered around its Leu8, Ile 44, and Val70 residues. The Vps27 UIM-1 is a helix with contiguous hydrophobic residues [182], Iota UBM-1 is characterized by a helix-turn-helix motif [183], and Rabex-5 binds to ubiquitin very similarly to Vps27 UIM-1, in a reverse orientation [184]. The Vps27 UIM-1 and Rabex-5 are helices whose amino acid sequence interacting with the hydrophobic patch of ubiquitin is constituted by alternating charged and nonpolar residues. The helix fold is such that the nonpolar residues face the receptor while the charged residues are oriented towards the solvent. Particularly, hydrophobic residues Leu262, Ile 263, Ala266, Ile267, and Leu271 of Vps27 UIM-1, and Ile51, Trp55, Leu57, Ala58, and Leu61 of Rabex-5are involved in hydrophobic ligand-protein 'contacts'. The ligand iota UBM1 of complex 2MBB contains less charged residues, with a nonpolar sequence, i.e., Leu78-Pro79-Val80, that is located at the level of its turn. The sequence is surrounded by other nonpolar residues facing the receptor, like Pro67, Val70, Val74, Phe75, Ile82 and Ile86.

Each of the Leu8, Ile44, and Val70 residues of ubiquitin belongs to a β-strand, i.e., β1, β3, and β5, respectively. Minimum distance maps between each ligand and its receptor (Figure 8.12), obtained from the analysis of the 100 ns all-atom MD trajectories described below, clearly show three regions of minimal distances involving those three strands. In complexes 1Q0W and 2MBB, about the whole sequence of the ligand closely interact with strands β1 (residues 1 to 8), β3 (residues 40 to 45), and β5 (residues 66 to 72) of ubiquitin.

MD simulations carried out with GROMACS 4.5.5 [185,186] and the Amber99SB FF [187] were first applied to model all-atom and RPCM structures, under conditions described in our last work, i.e., an equilibration stage of 40.1 ns followed by a 20 ns production stage [69]. All crystal water molecules

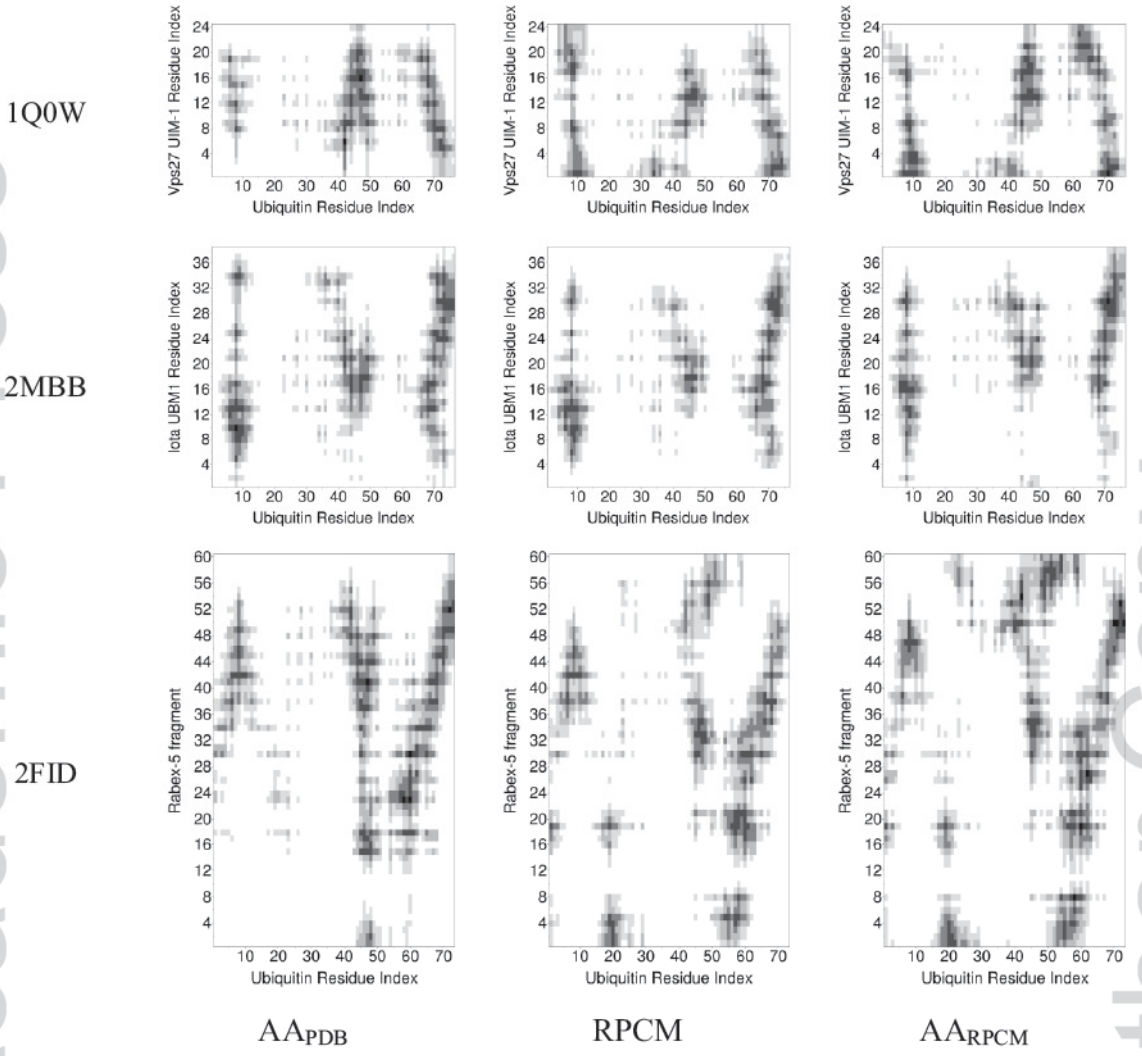


FIGURE 8.12 Ligand-receptor minimum distance maps calculated from the 100 ns production stages of the AA_{PDB} and AA_{RPCM} MD simulations and from the 20 ns production stage of the RPCM MD simulation. (Top) Vps27 UIM-1–Ubiquitin (PDB code: 1Q0W), (center) iota UBM1–Ubiquitin (PDB code: 2MBB), and (bottom) a bovine Rabex-5 fragment complexed with Ubiquitin (PDB code: 2FID). Scale goes from 0 to 1.5 nm (black to white) using a distance increment of 0.30 nm.

and the Zn ion of structure 2FID were removed. Structures were solvated in TIP4P-Ew water [188], and sodium ions were used to cancel the total electric charge. The final conformations of the 20 ns RPCM MD trajectories were considered as starting points for additional all-atom simulations, named here AA_{RPCM}, carried out with a 20 ns equilibration stage and a 100 ns production stage (50×10^6 steps with a time step of 0.002 ps) in the NPT ensemble at 1 bar and 300 K. Frames were saved every 20,000 steps. In addition, the initial all-atom MD trajectory was extended by a 100 ns long calculation. Thus, for each

of the three protein systems, three MD trajectories were analyzed, i.e., a 100 ns long all-atom MD started from the PDB structure of the complex (AA_{PDB}), a 20 ns long RPCM MD, and a 100 ns long all-atom MD started from the final RPCM frame (AA_{RPCM}). In addition, all-atom 100 ns long MD simulations of the unbound ligands in water were also carried out at 1 bar and 300 K.

The minimum ligand-ubiquitin distance maps established from the 20 ns RPCM and the 100 ns AA_{RPCM}MD trajectories show that the three β-strands of ubiquitin mentioned above are still involved in the interaction with the ligands like they are during the AA_{PDB} simulations (Figure 8.12). Particularly, RPCM and AA_{RPCM} maps of the iota UBM1 ligand are significantly similar to the corresponding AA_{PDB} map. It is due to structural changes that are weaker in complex 2 MBB than in the two other protein complexes, as illustrated by the last conformation of the MD trajectories (Figure 8.13). In the case of Vps27 UIM-1, the appearance of a turn during the RPCM-based simulation induces a break in the first region of the corresponding minimum distance map, i.e., below residue 15 of ubiquitin. The turn involves residues Ala266-Ile267-Glu268-Leu269 of the ligand (residues 12 to 15), and also appears in 100 ns all-atom MD trajectories of the unbound ligand of structure 1Q0W simulated in water (Figure 8.14). This is thus a weak point in the ligand sequence whose regularity is easily perturbed when the RPCM model is used. The all-atom MD simulation of the isolated 2MBB ligand preserves the turn involving residues Leu78-Pro79-Val80 (residues 17 to 19 in Figure 8.14), and two end chain bents appear at the level of residues Gly89-Lys90 (residues 28 and 29) and Gly69 (residue 8). Regarding the Rabex-5 ligand, the structural changes that occur during the AA_{RPCM} MD simulation of the complex lead to a minimum distance map that is similar to the one obtained from the 20 ns RPCM simulation (Figure 8.12). The initial helix structure of the unbound ligand of 2FID simulated at the all-atom level is now destructed at residues Ser36, Ile51-Glu52, and Glu65 (residues 23, 38–39, and 52 in Figure 8.14).

Very interestingly, there is a rather good match between the residues involved in non-helix regions of the ligands and those previously identified through a CWT analysis (Figure 8.6). Indeed, low absolute values of the wavelet coefficients actually correspond either to deconstructed regions of the initial a-helix structure of the three ligands, or turns, or characterize residues that precede preserved α-helix segments (Figure 8.14). It may involve that an α-propensity set of descriptors such as BLAM930101 [38] is useful in

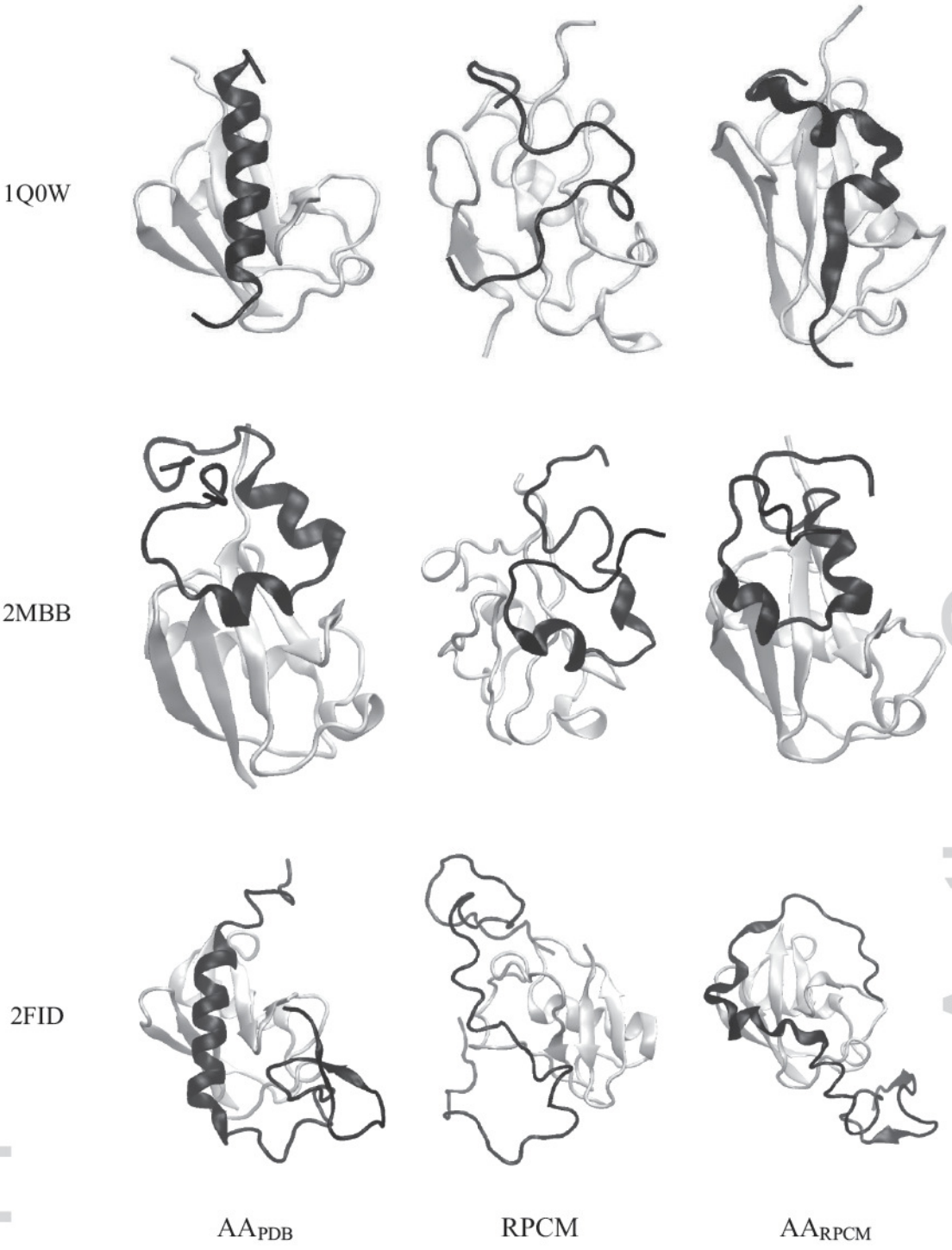


FIGURE 8.13 Last conformation of the three Ubiquitin complexes as obtained from the 100 ns production stages of the AA_{PDB} and AA_{RPCM} MD simulations and from the 20 ns RPCM simulation. (Top) Vps27 UIM-1–Ubiquitin (PDB code: 1Q0W), (center) iota UBM1–Ubiquitin (PDB code: 2MBB), and (bottom) a bovine Rabex-5 fragment complexed with Ubiquitin (PDB code: 2FID). The ligand is displayed in black.

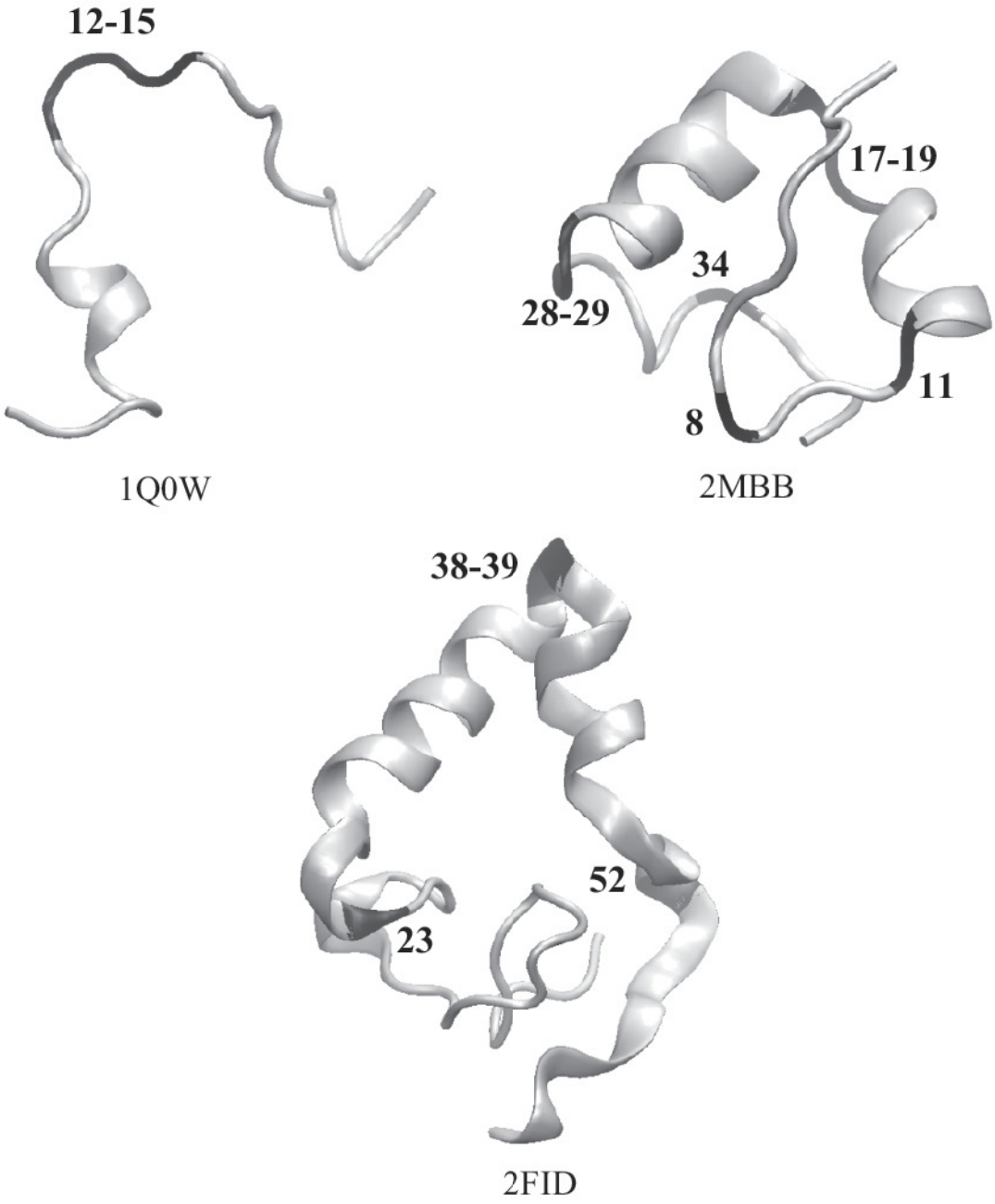


FIGURE 8.14 Last conformations of the three unbound Ubiquitin ligands simulated in water as obtained from 100 ns all-atom MD production stages. Residues occurring in destructed helix regions and turns are numbered and shown in black.

wavelet-based analysis to detect amino acid regions that are likely to lose their α-helical structure.

Figure 8.13 also illustrates that the AA_{PDB} conformations stay close to the PDB structure (Figure 8.1), except for 2FID whose long helix structure is partly deconstructed. In the complex 2FID, major structural changes appear

during the AA_{PDB} simulation due to the selection of a 1:1 ligand-ubiquitin complex while, in the crystal structure, the ligand is actually in close interaction with two ubiquitin molecules, the second one also interacting with a Zn ion. As mentioned by Chakrabartty et al., isolated helices derived from proteins are unstable due to the lack of side chain interactions [189], which initially occurred in the present case between the receptor and the ligand. A loss of secondary structure elements is also seen in RPCM conformations (Figure 8.13). The trend of 1Q0W and 2FID ligands to be deconstructed during RPCM simulations can be explained by the presence, in these two peptides, of numerous charged residues, i.e., 11 in 1Q0W and 19 in 2FID. Particularly, for 1Q0W and 2FID ligands, a cluster of five contiguous charged residues are present, at the level of the N-terminal and C-terminal ends, respectively. Thus, a modification in the point charge description is likely to more strongly affect their dynamical behavior than for the 2MBB ligand.

The return to all-atom interactions in AA_{RPCM} simulations allows to recover some regular secondary features in the structures. As emphasized earlier [67], RPCMs can lead to deformed conformations of the systems due to a lack of short-range electrostatic descriptions, which affect, notably, the existence of intra- and inter-molecular H-bonds. However, such conformations may appear to remain stable under all-atom MD simulation conditions. The RMSD versus the initially optimized PDB conformations of the complexes as well as the number of clusters detected during the production stages are presented in Figure 8.15 and Table 8.1, respectively. The low variation of the RMSD functions and the small number of clusters emphasize the stability of the AA_{RPCM} trajectories, thus providing clues that RPCM simulations can lead to the sampling of diverse and stable conformations. For complex 2FID, the AA_{RPCM} RMSD function is even lower than the AA_{PBD} one (Figure 8.15). This higher stability comes with an increased number of ligand-ubiquitin H-bonds, i.e., 15 rather than 11 (Table 8.1).

The number of clusters is determined using the approach called ‘Gromos’ method [190]. It consists, for each conformation in a trajectory, in counting the number of similar frames (called neighbors) considering a cut-off. The structure with the largest number of neighbors, and all its neighbors, are assumed to form a cluster, which is eliminated from the pool of already existing clusters. The procedure is repeated for the remaining structures in the pool. Cut-off values of 0.3 and 0.45 nm were selected to probe the receptor and ligand conformations, respectively. These values were chosen so as to keep small numbers of clusters for the ligand and ubiquitin when

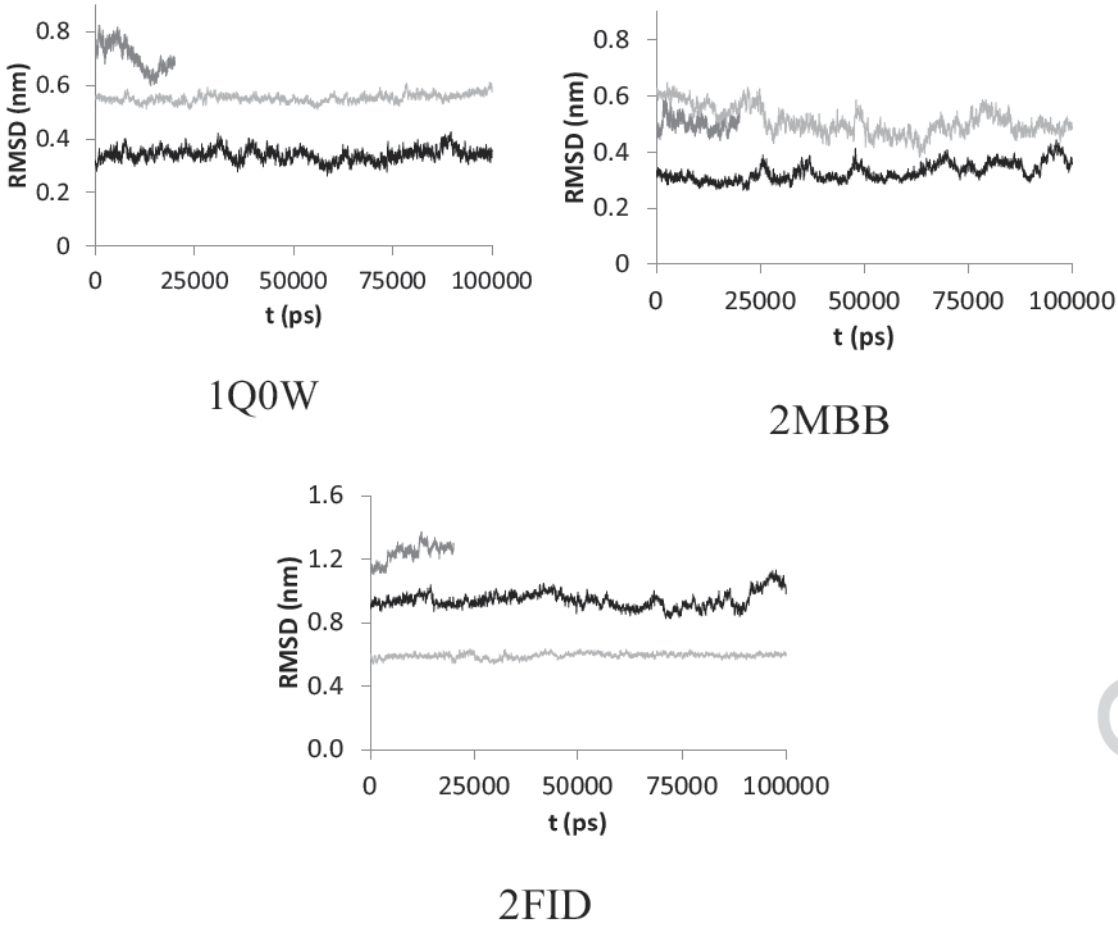


FIGURE 8.15 Time dependence of the protein RMSD calculated versus the initially optimized solvated structure from the 100 ns production stages of the AA_{PDB} and AA_{RPCM} MD simulations and from the 20 ns production stage of the RPCM MD simulation. Plain line = AA_{PDB}, gray line = RPCM, light gray line = AA_{RPCM}.

simulated at the AA_{PDB} level. Indeed, smaller cut-off values would involve additional clusters. Models RPCM lead to high numbers of clusters observed for the ligands, denoting a more flexible structure. The ligand iota UBM1 involved in complex 2MBB is an exception where the limited number of clusters reflects the high stability of the complex even at the RPCM level of representation. Despite a large number of AA_{RPCM} clusters, i.e., three, the good conformational conservation of the 2MBB ligand might be explained by the longer amino acid sequence of the ligand interacting with ubiquitin, i.e., 38 residues rather than 24 for 1Q0W. In all cases, the ubiquitin structure stays rather stable during all production stages.

A close examination of the H-bonds formed between the ligand and ubiquitin (Table 8.1) shows that the ligand forms less numerous H-bonds

TABLE 8.1 Properties (Mean and Standard Deviation Values) of the Solvent Layers of Thickness 0.35 nm Around the Protein Systems*

	AA_{PDB}	RPCM	AA_{RPCM}
MD production stage (ns)	100	20	100
1Q0W			
No. of molecules	1953 ± 46	2264 ± 64	2063 ± 65
No. of water-water H-bonds	436 ± 20	537 ± 21	464 ± 25
No. of protein-water H-bonds	282 ± 9	291 ± 10	312 ± 12
No. of ligand-Ubiquitin H-bonds	5 ± 1	1 ± 1	5 ± 2
No. of ligand/Ubiquitin clusters	1/1	13/1	1/1
2MBB			
No. of molecules	2084 ± 60	2444 ± 67	2155 ± 81
No. of water-water H-bonds	465 ± 25	582 ± 26	486 ± 30
No. of protein-water H-bonds	306 ± 10	294 ± 7	310 ± 13
No. of ligand-Ubiquitin H-bonds	5 ± 2	2 ± 1	3 ± 2
No. of ligand/Ubiquitin clusters	1/1	2/1	3/1
2FID			
No. of molecules	2522 ± 63	3047 ± 68	2579 ± 78
No. of water-water H-bonds	572 ± 26	737 ± 25	590 ± 28
No. of protein-water H-bonds	364 ± 10	375 ± 10	388 ± 15
No. of ligand-Ubiquitin H-bonds	11 ± 2	3 ± 2	15 ± 2
No. of ligand/Ubiquitin clusters	4/1	6/1	2/1

*The number of conformational clusters observed during the MD production stages is also given for the ligand and Ubiquitin.

at the RPCM level than at the all-atom level. For example, one H-bond rather than five is formed in structure 1Q0W, while only two and three are present in the 2MBB and 2FID complexes, respectively. One also observes that for complexes 1Q0W and 2FID, a high number of ligand-ubiquitin H-bonds are recovered at the AA_{RPCM} level versus their corresponding value obtained during the AA_{PDB} simulations. Values of 5 ± 2 and 15 ±2 rather than 5 ± 1 and 11 ± 2 are indeed obtained (Table 8.1). Contrarily, for structure 2MBB, a lower number of H-bonds at the AA_{RPCM} level, i.e., 3 ± 2 rather than 5 ± 2, is observed, but this does not affect the orientation of the ligand versus the receptor as conformations are rather similar (Figure 8.13).

The analysis of the protein-solvent interactions is also carried out by focusing on the hydrogen bonds. A water layer of thickness 0.35 nm was determined

around the protein complex and the mean numbers of water-water and water-protein H bonds were calculated considering that layer of solvent molecules. Analyses of the geometrical properties of the H-bonds were already reported previously [67, 68]. They showed that the usually expected first layer of water molecules is not structured any longer and intramolecular H-bonds strongly lose their orientation preference as reflected by the H-Donor-Acceptor angle values. The donor–acceptor distance distribution is however rather well preserved. Regarding the protein-water H-bonds, both distance and angle distributions stay similar to the all-atom results. From Table 8.1, one observes a larger number of water molecules in the layer close to the protein structure when the RPCM is used. For example, in structure 1Q0W, one gets 2264 molecules rather than 1953 and 2063 in the AA_{PDB} and AA_{RPCM} cases, respectively. It comes with an increase in the number of water-water H-bonds, i.e., 266 versus 217 and 233. Contrarily, the number of protein-water H-bonds is rather similar between the three charge models. For example, in structure 1Q0W, the number of such H-bonds is 282, 291, and 312, for trajectories AA_{PDB}, RPCM, and AA_{RPCM}, respectively. The AA_{RPCM} and AA_{PDB} trajectories behave similarly in terms of water-water H-bonds. It was actually shown previously that the use of a RPCM make the system very sensitive to the choice of the water force field [68]. The solvent is largely over-structuring, in opposition to what is known at the all-atom level [191, 192].

A statistical analysis of the potential energy terms calculated from the MD trajectories is reported in Table 8.2. Regarding the RPCM simulations, a post-processing calculation of the energy terms was carried out at the all-atom level. A good agreement is seen between the ligand-ubiquitin interaction energy terms calculated for the RPCM conformations using the RPCM and all-atom FFs. For example, values of –310.00 and –309.96 $kJ.mol^{-1}$ are obtained for the system 1Q0W. The ligand-ubiquitin energy contributions calculated at the all-atom level for the RPCM conformations are systematically higher than those obtained from the AA_{PDB} and AA_{RPCM} trajectories. Indeed, the latters involve additional stabilizing interactions like H-bonds. As an example, the system 1Q0W is characterized by intermolecular ubiquitin-ligand energy values of –309.96 $kJ.mol^{-1}$ rather than –474.22 and –571.81 $kJ.mol^{-1}$. The all-atom version of the RPCM intramolecular potential energy term is also always larger than the corresponding AA_{PDB} and AA_{RPCM} values. For example, one observes values of 20109.40, 18391.47, and 18984.40 $kJ.mol^{-1}$, for the RPCM, AA_{PDB}, and AA_{RPCM} simulations, respectively. It illustrates that a change in the point charge model involves short-range constraints

TABLE 8.2 All-Atom Mean Potential Energy Components ($kJ.mol^{-1}$) and Standard Deviation Calculated From the Production Stages of the MD Simulations**

MD production stage (ns)	AA_{PDB}	RPCM	AA_{RPCM}
	100	20	100
1Q0W			
Intramolecular ubiquitin and ligand	18391.47 ± 195.57	20109.40 ± 191.14* *16378.33 ± 180.78*	18984.40 ± 236.71
Intermolecular ligand-solvent	–4238.10 ± 200.47	–4826.66 ± 230.72* *–4893.84 ± 227.79*	–4521.04 ± 210.44
Intermolecular ubiquitin-solvent	–8913.81 ± 258.05	–9556.34 ± 259.89* *–9834.53 ± 270.38*	–9517.71 ± 345.16
Intermolecular ligand-ubiquitin	–474.22 ± 72.52	–309.96 ± 88.11* *–310.00 ± 90.95*	–517.81 ± 88.58
2MBB			
Intramolecular ubiquitin and ligand	22890.77 ± 199.86	24099.64 ± 234.47* *18311.47 ± 206.65*	22672.63 ± 245.78
Intermolecular ligand-solvent	–5427.15 ± 210.47	–5782.59 ± 195.94* *–5945.99 ± 197.60*	–5377.71 ± 263.47
Intermolecular ubiquitin-solvent	–8658.39 ± 266.12	–8694.14 ± 262.15* *–8987.46 ± 262.91*	–8788.56 ± 322.74
Intermolecular ligand-ubiquitin	–603.10 ± 102.54	–307.30 ± 66.47* *–304.33 ± 61.62*	–358.22 ± 92.73
2FID			
Intramolecular ubiquitin and ligand	25861.16 ± 226.01	28035.89 ± 208.74* *22115.24 ± 181.82*	26576.76 ± 306.60
Intermolecular ligand-solvent	–8826.55 ± 290.96	–9966.49 ± 241.66* *–10203.90 ± 244.34*	–9631.28 ± 293.92
Intermolecular ubiquitin-solvent	–7843.12± 237.08	–8373.19 ± 228.68* *–8655.84 ± 234.09*	–7833.00 ± 387.20
Intermolecular ligand-ubiquitin	–868.46 ± 93.99	–604.86 ± 137.12* *–602.52 ± 137.77*	–1176.00 ± 116.91

*Values obtained with the all-atom AMBER99SB FF applied to the conformations generated during the RPCM MD simulations.

**Long-range electrostatic contributions, calculated in the reciprocal space, are not considered in the reported values. Values in italics are energy terms calculated with the RPCM charges.

to the proteins. Those constraints are obviously less affecting the intermolecular energy terms. Table 8.2 also shows that the protein-solvent contributions are systematically lower for the RPCM models versus its all-atom counterpart. The case of the complex 1Q0W is again given here as an example, with values of –4893.84 and –9834.53 versus –4826.66 and –9556.34 kJ.mol^{-1}, for the RPCM and all-atom energy terms, respectively. It can be an explanation to the increased influence of the solvent when a RPCM is used [68].

An interesting point to mention is related to the stability of the AA_{RPCM} conformations, which may be characterized by more stabilizing ligand-solvent and ubiquitin-solvent interaction energy values than in the AA_{PDB} case. This is verified for complex 1Q0W, ant partly for systems 2MBB and 2FID where the ubiquitin-solvent and ligand-solvent energy values are more stabilizing than in the AA_{PDB} cases with –8788.56 and –9631.28 kJ.mol^{-1}, respectively. However, the ligand-solvent and ubiquitin-solvent counterparts are only slightly less stabilizing with –5377.71 and –7833.00 kJ.mol^{-1}, respectively. Also, in the cases of 1Q0W and 2FID, the intermolecular ligand-ubiquitin energy values calculated from the AA_{RPCM} trajectories are lower than in the AA_{PDB} cases. This is not observed at all for 2MBB, a nevertheless well-preserved complex, with a value of –358.22 versus –603.10 kJ.mol^{-1}.

8.4 CONCLUSIONS AND PERSPECTIVES

The current development of computer resources allows to study more and more complex systems over extended time scales. Consequently, more and more data are generated for storage and analysis. The need for simple models thus remains crucial to decrease the complexity of a problem, to get rid of useless data, or to reduce calculation time. Leveling and coarse-graining procedures remain widely used and still present vivid perspectives in the modeling of complex molecular systems and in the interpretation of experimental low-resolution data.

Two points of views can be adopted when using reduced molecular descriptions. Either one focuses on the generation of results that are similar to those obtained at a higher level of detail, but at a lower cost and with simpler algorithms, or one wishes to get results that differ from those obtained at a higher level of detail, thus providing new and/or different insights to a problem. In literature, various approaches are reported which, when applied to a same problem, may lead to different results.

Well-known methods such as spline approximation, Gaussian smoothing, wavelet multi-resolution analysis, and crystallography reconstruction, allow to level molecular properties. Discretization approaches such as vector quantization, critical point analysis, and coarse-graining procedures, generate models that replace a molecular property or representation by a limited number of data points. Leveling techniques are very common, e.g., in molecular graphics representations. Applications are found in many other research fields like global structure optimization, denoising of modeled or experimental data, molecular structure elucidation, similarity analysis, and molecular simulations.

We present selected results obtained from the implementation of smoothed electron density and charge density distribution functions in similarity of small molecules and molecular dynamics applications of proteins. Smoothing in similarity applications affects the number of possible solutions but also provides new or different solutions to problem. For example, it tends to align molecules in terms of global shape rather than in terms of atoms. On the other hand, the use of reduced point charge models of protein complexes favors ligand conformations that are not particularly stabilized at the all-atom level. Such conformations can however appear to be stable when returning to the all-atom description. A change in the level of detail of a protein complex is associated with a modification in the protein-solvent interactions. Thus, a focus on the influence of the solvent, especially in regards to the balance of protein-solvent interactions, and on the properties of protein systems as a function of the coarse-graining degree of the solvent, are planned as perspectives to the results reported in the present chapter.

ACKNOWLEDGMENTS

This research used resources of the 'Plateforme Technologique de Calcul Intensif (PTCI)' (http://www.ptci.unamur.be) located at the University of Namur, Belgium, which is supported by the F.R.S.-FNRS. The PTCI is member of the 'Consortium des Équipements de Calcul Intensif (CÉCI)' (http://www.ceci-hpc.be). The author gratefully acknowledges Daniel Vercauteren, Director of the 'Laboratory of Computational Physical Chemistry' at the University of Namur.

KEYWORDS

- charge density
- coarse graining
- denoising
- discretization
- electron density
- Gaussian convolution
- molecular dynamics
- molecular similarity
- multiresolution analysis
- point charge
- protein
- smoothing
- spline
- ubiquitin
- vector quantization
- wavelets

REFERENCES

1. Klasson, K. T., (2008). Construction of spline functions in spreadsheets to smooth experimental data. *Adv. Eng. Softw.*, *39*, 422–429.
2. Oberlin, D., Jr., & Scheraga, H. A., (1998). B-spline method for energy minimization in grid-based molecular mechanics calculations. *J. Comput. Chem.*, *19*, 71–85.
3. Humphrey, W., Dalk, A., & Schulten, K., (1996). VMD – Visual Molecular Dynamics. *J. Mol. Graphics*, *14*, 33–38.
4. Carson, M., (1987). Ribbon models of macromolecules. *J. Mol. Graphics*, *5*, 103–106.
5. Carson, M., (1996). Wavelets and molecular structure. *J. Comput. Aided Mol. Des.*, *10*, 273–283.
6. Bajaj, C. L., Pascucci, V., Shamir, A., Holt, R. J., & Netravali, A. N., (2003). Dynamic maintenance and visualization of molecular surfaces. *Discrete Appl. Math.*, *127*, 23–51.
7. Moré, J. J., & Wu, Z., (1997). Global continuation for distance geometry problems. *SIAM J. Optim.*, *7*, 814–836.
8. Kostrowicki, J., Piela, L., Cherayil, B. J., & Scheraga, H. A., (1991). Performance of the diffusion equation method in searches for optimum structures of clusters of Lennard-Jones atoms. *J. Phys. Chem.*, *95*, 4113–4119.

9. Wawak, R. J., Gibson, K. D., Liwo, A., & Scheraga, H. A., (1996). Theoretical prediction of a crystal structure. *Proc. Natl. Acad. Sci. USA*, *93*, 1743–1746.
10. Gironés, X., Amat, L., & Carbó-Dorca, R., (1998). A comparative study of isodensity surfaces using ab initio and ASA density functions. *J. Mol. Graph. Model.*, *16*, 190–196.
11. Tsirelson, V. G., Avilov, A. S., Abramov, Y. A., Belokoneva, E. L., & Kitaneh, R., (1998). Feil, D. X-ray and electron diffraction study of MgO. *Acta Crystallogr. B*, *54*, 8–17.
12. Tsirelson, V. G., Abramov, Y., Zavodnik, V., Stash, A., Belokoneva, E., & Stahn, J., (1998). Critical points in a crystal and procrystal. *Struct. Chem.*, *9*, 249–254.
13. Mitchell, A. S., & Spackman, M. A., (2000). Molecular surfaces from the promolecule: A comparison with Hartree-Fock ab initio electron density surfaces. *J. Comput. Chem.*, *21*, 933–942.
14. Gironés, X., Carbó-Dorca, R., & Mezey, P. G., (2001). Application of promolecular ASA densities to graphical representation of density functions of macromolecular systems. *J. Mol. Graph. Model.*, *19*, 343–348.
15. Downs, R. T., Gibbs, G. V., Boisen, M. B., Jr., & Rosso, K. M., (2002). A comparison of procrystal and ab initio model representations of the electron-density distributions of minerals. *Phys. Chem. Miner.*, *29*, 369–385.
16. Gironés, X., Amat, L., Carbó-Dorca, R., (2002). Modeling large macromolecular structures using promolecular densities. *J. Chem. Inf. Comput. Sci.*, *42*, 847–852.
17. Bultinck, P., Carbó-Dorca, R., (2003). Van Alsenoy Ch. Quality of approximate electron densities and internal consistency of molecular alignment algorithms in molecular quantum similarity. *J. Chem. Inf. Comput. Sci.*, *43*, 1208–1217.
18. Amat, L., & Carbó-Dorca, R., (1997). Quantum similarity measures under atomic shell approximation: First order density fitting using elementary Jacobi rotations. *J. Comput. Chem.*, *18*, 2023–2039.
19. Amat, L., & Carbó-Dorca, R., (2016). Quantum similarity measures under atomic shell approximation: First order density fitting using elementary Jacobi rotations, http://iqc.udg.es/cat/similarity/ASA/funcset.html (accessed May 18).
20. Allen, F. H., (2002). The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallogr. B*, *58*, 380–388.
21. Groom, C. R., Bruno, I. J., Lightfoot, M. P., & Ward, S. C., (2016). The Cambridge Structural Database, *Acta Crystallogr. B*, *72*, 171–179.
22. Hart, R. K., Pappu, R. V., & Ponder, J. W., (2000). Exploring the similarities between potential smoothing and simulated annealing. *J. Comput. Chem.*, *21*, 531–552.
23. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al., (2009). Gaussian 09, Revision E.01, Gaussian Inc., Wallingford, CT, USA.
24. Hall, S. R., du Boulay, D. J., & Olthof-Hazekamp, R., Eds., (2000). Xtal3.7 System, University of Western Australia. The source code is available at http://xtal.sourceforge.net/ (accessed May 18, 2016).
25. Leherte, L., (2004). Hierarchical analysis of promolecular full electron density distributions: Description of protein structure fragments. *Acta Crystallogr. D*, *60*, 1254–1265.
26. Walczak, B., & Massart, D. L., (1997). Wavelets – Something for analytical chemistry? *Trends Anal. Chem.*, *16*, 451–463.
27. Leung, A. K.-M., Chau, F.-T., & Gao, J.-B., (1998). A review on applications of wavelet transform techniques in chemical analysis: 1989–1997. *Chemom. Intell. Lab. Syst.*, *43*, 165–184.
28. Ehrentreich, F., (2002). Wavelet transform applications in analytical chemistry. *Anal. Bioanal. Chem.*, *372*, 115–121.

29. Dinç, E., & Baleanu, D., (2007). A review on the wavelet transform applications in analytical chemistry. In: *Mathematical Methods in Engineering*; Tas, K., Tenreiro Machado, J. A., Baleanu, D., Eds., Springer: Dordrecht, The Netherlands, pp. 265-284.
30. Chau, F.-T., & Leung, A. K.-M., (2000). Applications of wavelet transform in spectroscopic studies. In: *Wavelets in Chemistry*; vol.22; Walczak, B., Ed., Elsevier. Amsterdam, The Netherlands, pp. 241-261.
31. Teitelbaum, H., (2000). Application of wavelet analysis to physical chemistry. In: *Wavelets in Chemistry*; vol. 22; Walczak, B., Ed., Elsevier: Amsterdam, The Netherlands, pp. 263-289.
32. Liò, P., (2003). Wavelets in bioinformatics and computational biology: State of art and perspectives. *Bioinformatics*, *19*, 2–9.
33. Shao, X.-G., Leung, A. K.-M., & Chau, F.-T., (2003). Wavelet: A new trend in chemistry. *Acc. Chem. Res.*, *36*, 276–283.
34. Sundling, C. M., Sukumar, N., Zhang, H., Embrechts, M. J., & Breneman, C. M., (2006). Wavelets in chemistry and cheminformatics. *Rev. Comput. Chem.*, *22*, 295–329.
35. Alsberg, B. K., Woodward, A. M., & Kell, D. B., (1997). An introduction to wavelet transforms for chemometricians: A time-frequency approach. *Chemom. Intell. Lab. Syst.*, *37*, 215–239.
36. Jetter, K., Depczynski, U., Molt, K., & Niemöller, A., (2000). Principles and applications of wavelet transformation to chemometrics. *Anal. Chim. Acta*, *420*, 169–180.
37. Matlab and Statistics Toolbox, Release 2012b; The MathWorks, Inc.: Natick, MA, 2012.
38. Blaber, M., Zhang, X. J., & Matthews, B. W., (1993). Structural basis of amino acid alpha helix propensity. *Science*, *260*, 1637–1640.
39. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M., (2008). AAindex: Amino acid index database, progress report 2008. *Nucl. Acids Res.*, *36*, D202–D205.
40. Saha, I., Maulik, U., Bandyopadhyay, S., & Plewczynski, D., (2012). Fuzzy clustering of physicochemical and biochemical properties of amino acids. *Amino Acids 43*, 583–594.
41. Stollnitz, E. J., DeRose, T. D., & Salesin, D. H., (1995). Wavelets for computer graphics: A primer 1. *IEEE Comput. Graphics Appl.*, *15*(3), 76–84.
42. Stollnitz, E. J., DeRose, T. D., & Salesin, D. H., (1995). Wavelets for computer graphics: A primer 2, *IEEE Comput. Graphics Appl*., *15*(4), 75–85.
43. Mallat, S. G. A., (1989). theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.*, *11*, 674–693.
44. Jawerth, B., & Sweldens, W., (1993). An overview of wavelet based multiresolution analyses; Technical Report, University of South Carolina, USA. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.5864 (accessed May 18, 2016).
45. Nielsen, O. M., (1998). Wavelets in scientific computing, Ph.D. Dissertation, Technical University of Denmark, Lyngby, http://orbit.dtu.dk/fedora/objects/orbit:83353/datastreams/file_5265681/content (accessed May 18, 2016).
46. Daubechies, I., (1988). Orthonormal bases of compactly supported wavelets. *Commun. Pure Appl. Math.*, *41*, 909–996.
47. Main, P., & Wilson, J., (2000). Wavelet analysis of electron density maps. *Acta Crystallogr. D*, *56*, 618–624.
48. Starck, J.-L., Murtagh, F., & Bijaoui, A., (1998). *Image Processing and Data Analysis: The Multiscale Approach*; Cambridge University Press: Cambridge, United Kingdom, pp. 287.
49. González-Audícana, M., Otazu, X., Fors, O., & Seco, A., (2005). Comparison between Mallat's and the 'à trous' discrete wavelet transform based algorithms for

the fusion of multispectral and panchromatic images. *Int. J. Remote Sensing*, *26*, 595–614.
50. Shensa, M. J., (1992). The discrete wavelet transform: Wedding the à trous and the Mallat algorithms. *IEEE Trans. Signal Process.*, *40*, 2464–2482.
51. Voter, A. F., (1997). A method for accelerating the molecular dynamics simulation of infrequent events. *J. Chem. Phys*., *106*, 4665–4677.
52. Hamelberg, D., Mongan, J., & McCammon, J. A., (2004). Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys*., *120*, 11919–11929.
53. Perez, D., Uberuaga, B. P., Shim, Y., Amar, J. G., & Voth, A. F., (2009). Accelerated molecular dynamics methods: Introduction and recent developments. *Annu. Rep. Comput. Chem*., *5*, 79–98.
54. Leone, V., Marinelli, F., Carloni, P., & Parrinello, M., (2010). Targeting biomolecular flexibility with metadynamics. *Curr. Opin. Struct. Biol.*, *20*, 148–154.
55. Barducci, A., Bonomi, M., & Parrinello, M., (2011). *WIREs Comput. Mol. Sci*., *1*, 826–843.
56. Goedecker, S., (2004). Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys*., *120*, 9911–9917.
57. Wales, D. J., & Doye, J. P. K., (1997). Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A*, *101*, 5111–5116.
58. De-Alarcón, P. A., Pascual-Montano, A., Gupta, A., & Carazo, J. M., (2002). Modeling shape and topology of low-resolution density maps of biological macromolecules. *Biophys. J.*, *83*, 619–632.
59. Wriggers, W., Milligan, R. A., & McCammon, J. A., (1999). Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.*, *125*, 185–195.
60. Wriggers, W., & Birmanns, S., (2001). Using Situs for flexible and rigid-body fitting of multiresolution single-molecule data. *J. Struct. Biol*., *133*, 193–202.
61. Vorojbev, Y. N., (2010). Blind docking method combining search of low-resolution binding sites with ligand pose refinement by molecular dynamics-based global optimization. *J. Comput. Chem.*, *31*, 1080–1092.
62. Glick, M., Robinson, D. D., Grant, G. H., & Richards, W. G., (2002). Identification of ligand binding sites on proteins using a multi-scale approach. *J. Amer. Chem. Soc*., *124*, 2337–2344.
63. Glick, M., Grant, G. H., & Richards, W. G., (2002). Docking of flexible molecules using multiscale ligand representations. *J. Med. Chem*., *45*, 4639–4646.
64. Leherte, L., Dury, L., & Vercauteren D. P., (2003). Structural identification of local maxima in low-resolution promolecular electron density distributions. *J. Phys. Chem. A*, *107*, 9875–9886.
65. Leung, Y., Zhang, J.-S., & Xu, Z.-B., (2000). Clustering by scale-space filtering. *IEEE Trans. Pattern Anal. Mach. Intell.*, *22*, 1396–1410.
66. Leherte, L., & Vercauteren, D. P., (2009). Coarse point charge models for proteins from smoothed molecular electrostatic potentials. *J. Chem. Theory Comput.*, *5*, 3279–3298.
67. Leherte, L., & Vercauteren, D. P., (2014). Evaluation of reduced point charge models of proteins through Molecular Dynamics simulations: Application to the Vps27 UIM-1–ubiquitin complex. *J. Mol. Graphics Model.*, *47*, 44–61.

68. Leherte, L., & Vercauteren, D. P., (2014). Comparison of reduced point charge models of proteins: Molecular dynamics simulations of ubiquitin. *Sci. China Chem.*, *57*, 1340–1354.
69. Leherte, L., (2016). Reduced point charge models of proteins: Assessment based on molecular dynamics simulations. *Mol. Simul.*, *42*, 289–304.
70. Johnson, C. K., (1977). Orcrit. The Oak Ridge Critical Point Network Program; Technical report; Oak Ridge National Laboratory: Oak Ridge, TN.
71. Leherte, L., & Allen, F. H., (1994). Shape information from a critical point analysis of calculated electron density maps: Application to DNA-drug Systems. *J. Comput. Aided Mol. Des.*, *8*, 257–272.
72. Becue, A., Meurice, N., Leherte, L., & Vercauteren, D. P., (2003). Description of protein-DNA complexes in terms of electron density topological features. *Acta Crystallogr. D*, *59*, 2150–2162.
73. Becue, A., Meurice, N., Leherte, L., & Vercauteren, D. P., (2004). Evaluation of the protein solvent-accessible surface using reduced representations in terms of critical points of the electron density. *J. Comput. Chem.*, *25*, 1117–1126.
74. Leherte, L., (2001)Applications of multiresolution analyses to electron density maps of small molecules: Critical point representations for molecular superposition. *J. Math. Chem.*, *29*, 47–83.
75. Burton, J., Meurice, N., Leherte, L., & Vercauteren, D. P., (2008). Can descriptors of the electron density distribution help to distinguish functional groups. *J. Chem. Inf. Model.*, *48*, 1974–1983.
76. Troyer, J. M., & Cohen, F. E., (1990). Simplified models for understanding and predicting protein structure, *Rev. Comput. Chem.*, *2*, 57–80.
77. Piela, L., (1998). Search for the most stable structures on potential energy surfaces. *Collect. Czech. Chem. Commun.*, *63*, 1368–1380.
78. Pappu, R. V., Hart, R. K., & Ponder, J. W., (1998). Analysis and applications of potential energy smoothing and search methods for global optimization. *J. Phys. Chem. B*, *102*, 9725–9742.
79. TINKER – Software Tools for Molecular Design, Jay Ponder Lab, Department of Chemistry, Washington University, Saint Louis (MI) USA, (2015). http://dasher.wustl.edu/tinker/ (accessed 30 June 2016).
80. Schelstraete, S., Schepens, W., & Verschelde, H., (1999). Energy minimization by smoothing techniques: A survey. In: *Molecular Dynamics: From Classical to Quantum Methods*; vol. 7; Balbuena, P. B., Seminario. J. M., Eds., Elsevier: Amsterdam, The Netherlands, pp. 129-185.
81. Grossfield, A., & Ponder, J. W., (2016). Global optimization via a modified potential smoothing kernel; CCB Report 2002-01; Washington University School of Medicine: St Louis, MO, (2002). http://dasher.wustl.edu/ponder/papers/ccb-report-2002-01.pdf (accessed May 18).
82. Goldstein, M., Fredj, E., & Gerber, R. B., (2011). A new hybrid algorithm for finding the lowest minima of potential surfaces: Approach and application to peptides. *J. Comput. Chem.*, *32*, 1785–1800.
83. Shao, C.-S., Byrd, R., Eshow, E., & Schnabel, R. B., (2011). Global optimization for molecular clusters applied to molecular structure determination. *Oper. Res. Lett.*, *39*, 461–465.
84. Moré, J. J., & Wu, Z., (1999). Distance geometry optimization for protein structures. *J. Global Optim.*, *15*, 219–234.

85. Liberti, L., Lavor, C., & Macula, N., (2009). Double variable neighborhood search with smoothing for the molecular distance geometry problem. *J. Global Optim*., *43*, 207–218.
86. Souza, M., Xavier, A. E., Lavor, C., & Maculan, N., (2011). Hyperbolic smoothing and penalty techniques applied to molecular structure determination. *Oper. Res. Lett*., *39*, 461–465.
87. Diller, D. J., Verlinde, & Ch, M. L. J., (1999). A critical evaluation of several optimization algorithms for the purpose of molecular docking. *J. Comput. Chem*., *20*, 1740–1751.
88. Cai, C., Gong, J., Liu, X., & Gao, D., Li, H., (2013). Molecular similarity: Methods and performances. *Chin. J. Chem*., *31*, 1123–1132.
89. Duncan, B. S., & Olson, A. J., (1993). Shape analysis of molecular surfaces. *Biopolymers*, *33*, 231–238.
90. Maggiora, G. M., Rohrer, D. C., & Mestres, J., (2001). Comparing protein structures: A Gaussian-based approach to the three-dimensional structural similarity of proteins. *J. Mol. Graph. Model.*, *19*, 168–178.
91. Leherte, L., (2006). Similarity measures based on Gaussian-type promolecular electron density models: Alignment of small rigid molecules. *J. Comput. Chem*., *27*, 1800–1816.
92. Leherte, L., & Vercauteren, D. P., (2012). Smoothed Gaussian molecular fields: An evaluation of molecular alignment problems, *Theor. Chem. Acc*., *131*, 1259/1–1259/16.
93. Maggiora, G. M., & Shanmugasundaram, V., (2004). Molecular similarity measures. In: *Chemoinformatics: concepts, methods, and tools for drug discovery*, vol. 275 of series: *Methods in Molecular Biology*; Bajorath, J., Ed., Humana Press: Totowa, NJ, USA, pp. 1-50.
94. Bultinck, P., Gironés, X., & Carbó-Dorca, R., (2005). Molecular quantum similarity: Theory and applications. *Rev. Comput. Chem*., *21*, 127–207.
95. Carbó-Dorca, R., Besalú, E., & Mercado, L. D., (2011). Communications on quantum similarity, Part 3: A geometric-quantum similarity molecular superposition algorithm. *J. Comput. Chem*., *32* 582–599.
96. Leherte, L., Meurice, N., & Vercauteren, D. P., (2005). Influence of conformation on the representation of small flexible molecules at low resolution: Alignment of endothiapepsin ligands. *J. Comput. Aided Mol. Des.*, *19*, 525–549.
97. Carbó, R., Leyda, L., & Arnau, M., (1980). How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem*., *17*, 1185–1189.
98. Carbó, R., & Calabuig, B., (1992). Molecular quantum similarity measures and N-dimensional representation of quantum objects. I. Theoretical foundations. *Int. J. Quantum Chem*., *42*, 1681–1693.
99. Carbó, R., & Besalú, E., (1995). Theoretical foundations of quantum molecular similarity. In: *Molecular Similarity and Reactivity – From Quantum Chemical to Phenomenological Approaches*; vol. 14 of series: *Understanding Chemical Reactivity*; Carbó-Dorca, R., Ed., Kluwer: Dordrecht, The Netherlands, pp. 3–28.
100. Carbó-Dorca, R., & Besalú, E., (1998). A general survey of molecular quantum similarity. *J. Mol. Struct. (Theochem)*, *451*, 11–23.
101. Robert, D., & Carbó-Dorca, R., (1998). A formal comparison between molecular quantum similarity measures and indices. *J. Chem. Inf. Comput. Sci*., *38*, 469–475.

102. Maggiora, G. M., Petke, J. D., & Mestres, J., (2002). A general analysis of field-based molecular similarity indices. *J. Math. Chem*., *31*, 251–270.
103. Monev, V., (2004). Introduction to similarity searching in chemistry. *Commun. Math. Comput. Chem*., *51*, 7–38.
104. Constans, P., & Carbó, R., (1995). Atomic Shell Approximation: Electron density fitting algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.*, *35*, 1046–1053.
105. Constans, P., Amat, L., & Fradera, X., & Carbó-Dorca, R., (1996). Quantum molecular similarity measures (QMSM) and the atomic shell approximation (ASA). *Adv. Mol. Simil*., *1*, 187–211.
106. Amat, L., & Carbó-Dorca, R., (1999). Fitted electronic density functions from H to Rn for use in quantum similarity measures: cis-diammine-dichloroplatinum(II) complex as an application example. *J. Comput. Chem., 20*, 911–920.
107. Amat, L., & Carbó-Dorca, R., (2000). Molecular electronic density fitting using elemantary Jacobi rotations under atomic shell approximation. *J. Chem. Inf. Comput. Sci*., *40*, 1188–1198.
108. Constans, P., Amat, L., & Carbó-Dorca, R., (1997). Toward a global maximization of the molecular similarity function: Superposition of two molecules. *J. Comput. Chem*., *18*, 826–846.
109. Ritchie, D. W., & Kemp, G. J. L., (1999). Fast computation, rotation and comparison of low resolution spherical harmonic molecular surfaces. *J. Comput. Chem*., *20*, 383–395.
110. Hakkoymaz, H., Kieslich, Ch.A., Gorham, R. D., Jr., Gunopulos, D., & Morikis, D., (2011). Electrostatic similarity determination using multiresolution analysis. *Mol. Inf*., *30*, 733–746.
111. Beck, M. E., & Schindler, M., (2009). Quantitative structure-activity relations based on quantum theory and wavelet transformations. *Chem. Phys*., *356*, 121–130.
112. Martin, R. L., Gardiner, E. J., Senger, S., & Gillet, V. J., (2012). Compression of molecular interaction fields using wavelet thumbnails: Application to molecular alignment. *J. Chem. Inf. Model.*, *52*, 757–769.
113. Crippen, G. M., (2003). Series approximation of protein structure and constructing conformation space. *Polymer*, *44*, 4373–4379.
114. Fischer, P., Baudoux, G., & Wouters, J., (2003). Wavpred: A wavelet-based algorithm for the prediction of transmembrane proteins. *Comm. Math. Sci.*, *1*, 44–56.
115. Qiu, J., Liang, R., Zou, X., & Mo, J., (2003). Prediction of protein secondary structure based on continuous wavelet transform. *Talanta*, *61*, 285–293.
116. Vannuci, M., & Liò, P., (2001). Non-decimated wavelet analysis of biological sequences: Applications to protein structure and genomics. *Indian J. Stat.*, *63*, 218–233.
117. de Trad, C. H., Fang, Q., & Cosic, I., (2002). Protein sequence comparison based on the wavelet transform approach. *Prot. Eng.*, *15*, 193–203.
118. Sabarish, R. A., & Thomas, T., (2011). A frequency domain approach to protein sequence similarity analysis and functional classification. *Signal Image Process*, *2*, 36–48.
119. Chua, G.-H., Krishnan, A., Li, K.-B., & Tomita, M., (2006). Multiresolution analysis uncovers hidden conservation of properties in structurally and functionally similar proteins. *J. Bioinform. Comput. Biol.*, *4*, 1245–1267.
120. Wen, Z.-N., Wang, K.-L., Li, M.-L., Nie, F.-S., & Yang, Y., (2005). Analyzing functional similarity of protein sequences with discrete wavelet transform. *Comput. Biol. Chem.*, *29*, 220–228.

121. Krishnan, A., Li, K.-B., & Isaac, P., (2004). Rapid detection of conserved regions in protein sequences using wavelets. *In Silico Biol.*, *4*, 133–148.
122. Tsonis, A. A., Kumar, P., Elsner, J. B., & Tsonis, P. A., (1996). Wavelet analysis of DNA sequences. *Phys. Rev. E*, *53*, 1828–1834.
123. Machado, J.-A., Costa, A., & Quelhas, M. D., (2011). Wavelet analysis of human DNA. *Genomics*, *98*, 155–163.
124. Mena-Chalco, J., Carrer, H., Zana, Y., & Cesar, R. M., Jr., (2008). Identification of protein coding regions using the modified Gabor-wavelet transform. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, *5*, 198–207.
125. Barclay, V. J., & Bonner, R. F., (1997). Application of wavelet transforms to experimental spectra: Smoothing, denoising, and data set compression. *Anal. Chem.*, *69*, 78–90.
126. Pilard, M., & Epelboin, Y., (1998). Multiresolution analysis for the restoration of noisy X-ray topographs. *J. Appl. Crystallogr.*, *31*, 36–46.
127. Ergen, B., (2013). Comparison of wavelet types and thresholding methods on wavelet based denoising of heart sounds. *J. Signal Inf. Process*, *4*, 164–167.
128. Jeena, J., Salice, P., & Neetha, J., (2013). Denoising using soft thresholding, *Int. J. Adv. Res. Elec. Electron. Instrum. Eng.*, *2*, 1027–1032.
129. Chen, J.-S., Teng, H., & Nakano, A., (2007). Wavelet-based multi-scale coarse graining approach for DNA molecules. *Finite Elem. Anal. Des.*, *43*, 346–360.
130. Vakser, I. A., Matar, O. G., & Lam, C. F., (1999). A systematic study of low-resolution recognition in protein protein complexes. *Proc. Natl. Acad. Sci. USA*, *96*, 8477–8482.
131. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., & Vakser, I. A., (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA*, *89*, 2195–2199.
132. Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., et al., (2004). A structural perspective on protein-protein interactions. *Curr. Op. Struct. Biol.*, *14*, 313–324.
133. Vakser, I. A., (1996). Low-resolution docking: Prediction of complexes for underdetermined structures. *Biopolymers*, *39*, 455–464.
134. Tovchigrechko, A., Wells, Ch. A., & Vakser, I. A., (2002). Docking of protein models. *Prot. Sci.*, *11*, 1888–1896.
135. Nittinger, E., Schneider, N., Lange, G., & Rarey, M., (2015). Evidence of water molecules: A statistical evaluation of water molecules based on electron density. *J. Chem. Inf. Model.*, *55*, 771–783.
136. Emperador, A., Carrillo, O., Rueda, M., & Orozco, M., (2008). Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. *Biophys. J.*, *95*, 2127–2138.
137. Hinsen, K., (2008). Structural flexibility in proteins: Impact of the crystal environment. *Bioinform.*, *24*, 521–528.
138. Moritsugu, K., & Smith, J. C., (2008). REACH Coarse-grained biomolecular simulation: Transferability between different protein structural classes. *Biophys. J.*, *95*, 1639–1648.
139. Hills, R. D., Jr., Lu, L., & Voth, G. A., (2010). Multiscale coarse-graining of the protein energy landscape. *PLoS Comput. Biol.*, *6*, e1000827/1–e1000827/12.
140. Spijker, P., van Hoof, B., Debertrand, M., Markvoort, A. J., Vaidehi, N., & Hilbers, A. J., (2010). Coarse-grained molecular dynamics simulations of transmembrane protein-lipid systems. *Int. J. Mol. Sci.*, *11*, 2393–2420.
141. Doruker, P., Jernigan, R. L., & Bahar, I., (2002). Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J. Comput. Chem.*, *23*, 119–127.

142. Darré, L., Machado, M. R., Brandner, A. F., González, H. C., Ferreira, S., & Pantano, S., (2014). SIRAH: A structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. *J. Chem. Theory. Comput.*, *11*,723–739.
143. Basdevant, N., Ha-Duong, T., & Borgis, D., (2007). A coarse-grained protein-protein potential derived from an all-atom force field. *J. Phys. Chem. B*, *111*, 9390–9399.
144. Gohlke, H., & Thorpe, M. F., (2006). A natural coarse-graining for simulating large biomolecular motion. *Biophys. J.*, *91*, 2115–2120.
145. Arkhipov, A., Freddolino, P. L., & Schulten, K., (2006). Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure*, *14*, 1767–1777.
146. Arkhipov, A., Yin, Y., & Schulten, K., (2008). Four-scale description of membrane sculpting by BAR domains. *Biophys. J.*, *95*, 2806–2821.
147. Chng, Ch.-P., & Yang, L.-W., (2008). Coarse-grained models reveal functional dynamics – II. Molecular dynamics simulation at the coarse-grained level – Theories and biological applications. *Bioinform. Biol. Insights*, *2*, 171–185.
148. Yang, L.-W., & Chng, Ch.-P., (2008). Coarse-grained models reveal functional dynamics – I. Elastic network models – Theories, comparisons and perspectives. *Bioinf. Biol. Insights*, *2*, 25–45.
149. Kamerlin, S. C. L., Vicatos, S., Dryga, A., & Warshel, A., (2011). Coarse-grained (multiscale) simulations in studies of biophysical and chemical systems. *Annu. Rev. Phys. Chem*., *62*, 41–64.
150. Naganathan, A. N., (2013). Coarse-grained models of protein folding as detailed tools to connect with experiments. *WIREs Comput. Mol. Sci.*, *3*, 504–514.
151. Baaden, M., & Marrink, S. J., (2013). Coarse-grain modelling of protein-protein interactions. *Curr. Opin. Struct. Biol.*, *23*, 878–886.
152. Brini, E., Algaer, E. A., Ganguly, P., Li, C., Rodríguez-Ropero, F., & van der Vegt, N. F.A., (2013). Systematic coarse-graining methods for soft matter simulations – A review. *Soft Matter*, *9*, 2108–2119.
153. Meier, K., Choutko, A., Dolenc, J., Eichenberger, A. P., Riniker, S., & van Gunsteren, W. F., (2013). Multi-resolution simulation of biomolecular systems: A review of methodological issues. *Angew. Chem. Int. Ed.*, *52*, 2820–2834.
154. Saunders, M., & Voth, G. A., (2013). Coarse-graining methods for computational biology. *Annu. Rev. Biophys.*, *42*, 73–93.
155. Paramonov, L., & Yaliraki, S. N., (2005). The directional contact distance of two ellipsoids: Coarse-grained potentials for anisotropic interactions. *J. Chem. Phys.*, *123*, 194111/1–194111/11.
156. Izvekov, S., & Voth, G. A., (2005). A Multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B*, *109*, 2469–2473.
157. Liu, P., Izvekov, S., & Voth, G. A., (2007). Multiscale coarse-graining of monosaccharides. *J. Phys. Chem. B*, *111*, 11566–11575.
158. Carbone, P., Varnazeh, H. A.K., Chen, X., & Müller-Plathe, F., (2008). Transferability of coarse-grained force fields: The polymer case. *J. Chem. Phys.*, *128*, 064904/1–064904/11.
159. Naômé, A., Laaksonen, A., & Vercauteren, D. P., (2014). A solvent-mediated coarse-grained model of DNA derived with the systematic Newton inversion method. *J. Chem. Theory Comput.*, *10*, 3541–3549.
160. Kondrashov, D. A., Cui, Q., & Phillips, G. N., Jr., (2006). Optimization and evaluation of a coarse-grained model of protein motion using X-ray crystal data. *Biophys. J.*, *91*, 2760–2767.

161. Lyman, E., Pfaendtner, J., & Voth, G. A., (2008). Systematic multiscale parametrization of heterogeneous elastic network models of proteins. *Biophys. J.*, *95*, 4183–4192.
162. Orellana, L., Rueda, M., Ferrer-Costa, C., Lopez-Blanco, J. R., Chacón, P., & Orozca, M., (2010). Approaching elastic network models to molecular dynamics flexibility. *J. Chem. Theory Comput.*, *6*, 2910–2923.
163. Lyubartsev, A. P., & Laaksonen, A., (1995). Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Phys. Rev. E*, *52*, 3730–3737.
164. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., & de Vries, A. H., (2007). The MARTINI forcefield: Coarse-grained model for biomolecular simulations. *J. Phys. Chem. B*, *111*, 7812–7824.
165. Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., & Marrink, S. J., (2008). The MARTINI coarse-grained forcefield: Extension to proteins. *J. Chem. Theory Comput.*, *4*, 819–834.
166. Liwo, A., Czaplewski, C., Oldziej, S., Rojas, A. V., Kazmierkiewicz, R., Makowski, M., et al., (2009). In Coarse-graining of condensed phase and biomolecular systems; Voth, G. A., Ed., CRC Press: Boca Raton, FL.
167. Fujitsuka, Y., Takada, S., Luthey-Schulten, Z., & Wolynes, P. G., (2004). Optimizing physical energy functions for protein folding. *Proteins*, *54*, 88–103.
168. Hori, N., Chikenji, G., Berry, R. S., & Takada, S., (2009). Folding energy landscape and network dynamics of small globular proteins. *Proc. Natl. Acad. Sci. USA*, *106*, 73–78.
169. Clementi, C., (2008). Coarse-grained models of protein folding: Toy models or predictive tools? *Curr. Opin. Struct. Biol.*, *18*, 10–15.
170. Sherwood, P., Brooks, B. R., & Sansom, M. S. P., (2008). Multiscale methods for macromolecular simulations. *Curr. Opin. Struct. Biol.*, *18*, 630–640.
171. He, N., Liu, Y., & Zhang, X., (2016). Molecular dynamics - Smoothed molecular dynamics (MD-SMD) adaptive coupling method with seamless transition. *Int. J. Numer. Meth. Engng*. doi:10.1002/nme.5224.
172. Cherfils, J., Duquerroy, S., & Janin, J., (1991). Protein–protein recognition analyzed by docking simulation. *Proteins*, *11*, 271–280.
173. Sternberg, M. J. E., & Gabb, H. A., & Jackson, R. M., (1998). Predictive docking of protein–protein and protein–DNA complexes. *Curr. Opin. Struct. Biol.*, *8*, 250–256.
174. Butzlaff, M., Dahmen, W., Diekmann, S., Dress, A., Schmitt, E., & von Kitzing, E., (1994). A hierarchical approach to force field calculations through spline approximations. *J. Math. Chem.*, *15*, 77–92.
175. von Kitzing, E., & Schmitt, E., (1995). Configurational space of biological macromolecules as seen by semiempirical force fields: Inherent problems for molecular design and strategies to solve them by means of hierarchical force fields. *J. Mol. Struct. (Theochem)*, *336*, 245–259.
176. Olson, W. K., (1996). Simulating DNA at low resolution. *Curr. Opin. Struct. Biol.*, *6*, 242–256.
177. Dury, L., Latour, Th., Leherte, L., Barberis, F., & Vercauteren, D. P., (2001). A new graph descriptor for molecules containing cycles. Application as screening criterion for searching molecular structures within large databases of organic compounds. *J. Chem. Inf. Comput. Sci.*, *41*, 1437–1445.
178. Fischer, J. R., Lessel, U., & Rarey, M., (2011). Improving similarity-driven library design: Customized matching and regioselective feature trees. *J. Chem. Inf. Model.*, *51*, 2156–2163.

179. Birchall, K., & Gillet, V. J., (2011). Reduced Graphs and Their Applications in Chemoinformatics. In: *Cheminformatics and Computational Chemical Biology*, vol.672 of series: *Methods in Molecular Biology*; Bajorath, J., Ed., Humana Press: New York, NY, USA, pp. 197-212.
180. Jia, Z., & Chen, J., (2016). Necessity of high-resolution for coarse-grained modeling of flexible proteins. *J. Comput. Chem.*, *37*, 1725–1733.
181. Cao, Z., & Voth, G. A., (2015). The multiscale coarse-graining method. XI. Accurate interaction based on the centers of charge of coarse-grained sites. *J. Chem. Phys.*, *143*, 243116/1–243116/11.
182. Swanson, K. A., Kang, R. S., Stamenova, S. D., Hicke, L., & Radhakrishnan, I., (2003). Solution structure of Vps27 UIM-ubiquitin complex important for endosomal sorting and receptor downregulation. *EMBO J.*, *22*, 4597–4606.
183. Wang, S., & Zhou, P., (2014). Sparsely-sampled, high-resolution 4D omit spectra for detection and assignment of intermolecular NOEs of protein complexes. *J. Biomol. NMR*, *59*, 51–56.
184. Lee, S., Tsai, Y. C., Mattera, R., Smith, W. J., Kostelansky, M. S., Weissman, A. M., et al., (2006). Structural basis for ubiquitin recognition and autoubiquitination by Rabex-5. *Nat. Struct. Mol. Biol.*, *13*, 264–271.
185. Hess, B., Kutzner, C., van der Spoel, D., & Lindahl, E., (2008). GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, *4*, 435–447.
186. Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., et al., (2013). GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, *29*, 845–854.
187. Showalter, S. A., & Brüschweiler, R., (2007). Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: Application to the AMBER99SB force field. *J. Chem. Theory Comput.*, *3*, 961–975.
188. Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., & Head-Gordon, T., (2004). Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.*, *120*, 9665–9678.
189. Chakrabartty, A., Kortemme, T., & Baldwinn, R. L., (1994). Helix propensity of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Prot. Sci.*, *3*, 843–852.
190. Daura, X., Gademann, K., Jaun, B., Seebach, D., Gunsteren, W. F., & Mark, A. E., (1999). Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Ed.*, *38*, 236–240.
191. Best, R. B., Zheng, W., & Mittal, J., (2014). Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.*, *10*, 5113–5124.
192. Henriques, J., Gragnell, C., & Skepö, M., (2015). Molecular Dynamics simulations of intrinsically disordered proteins: Force field evaluation and comparison with experiment. *J. Chem. Theory Comput.*, *11*, 3420–3431.